Optimal model selection in density estimation

Matthieu Lerasle

Optimal model selection in density estimation

Matthieu Lerasle*

Abstract

We build penalized least-squares estimators using the slope heuristic and resampling penalties. We prove oracle inequalities for the selected estimator with leading constant asymptotically equal to 1. We compare the practical performances of these methods in a short simulation study.

Key words: Density estimation, optimal model selection, resampling methods, slope heuristic.
2000 Mathematics Subject Classification: 62G07, 62G09.

1 Introduction

The aim of model selection is to construct data-driven criteria to select a model among a given list. The history of statistical model selection goes back at least to Akaike [1], [2] and Mallows [18]. They proposed to select among a collection of parametric models the one which minimizes an empirical loss plus some penalty term proportional to the dimension of the model. Birgé & Massart [8] and Barron, Birgé & Massart [6] generalized this approach, making in particular the link between model selection and adaptive estimation. They proved that previous methods, in particular cross-validation (see Rudemo [20]) and hard thresholding (see Donoho *et.al.* [12]) can be viewed as penalization methods. More recently, Birgé & Massart [9], Arlot & Massart [5] and Arlot [4], (see also [3]) arised the problem of optimal efficient model selection. Basically, the aim is to select an estimator satisfying an oracle inequality with leading constant asymptotically equal to 1. They obtained such procedures thanks to a sharp estimator of the ideal penalty pen_{id}. We will be interested in two natural ideas, that are used in practice to evaluate pen_{id} and proved to be efficient in other frameworks. The first one is the slope heuristic. It was introduced in Birgé & Massart [9] in Gaussian regression and developed in Arlot & Massart [5] in a M-estimation framework. It allows to optimize the choice of a leading constant in the penalty term, provided that we know the shape of pen_{id}. The other one is Efron's resampling heuristic. The basic idea comes from Efron [14] and was used by Fromont [15] in the classification framework. Then, Arlot [4] made the link with ideal penalties and developed the general procedure. Up to our knowledge, these methods have only been theoretically validated in regression frameworks. We propose here to prove their efficiency in density estimation. Let us now explain more precisely our context.

1.1 Least-squares estimators

In this paper, we define and study efficient penalized least-squares estimators in the density estimation framework when the error is measured with the L^2-loss. We observe n

*Institut de Mathématiques (UMR 5219), INSA de Toulouse, Université de Toulouse, France

i.i.d random variables $X_1, ..., X_n$, defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, valued in a measurable space $(\mathbb{X}, \mathcal{X})$, with common law P. We assume that a measure μ on $(\mathbb{X}, \mathcal{X})$ is given and we denote by $L^2(\mu)$ the Hilbert space of square integrable real valued functions defined on $\mathbb{X}$. $L^2(\mu)$ is endowed with its classical scalar product, defined for all t, t' in $L^2(\mu)$ by

$$< t, t' >= \int_{\mathbb{X}} t(x)t'(x)d\mu(x)$$

and the associated L^2-norm $\|.\|$, defined for all t in $L^2(\mu)$ by $\|t\| = \sqrt{< t, t >}$. The parameter of interest is the density s of P with respect to μ, we assume that it belongs to $L^2(\mu)$. The risk of an estimator $\hat{s}$ of s is measured with the L^2-loss, that is $\|s - \hat{s}\|^2$, which is random when $\hat{s}$ is.
s minimizes the integrated quadratic contrast $PQ(t)$, where $Q : L^2(\mu) \to L^1(P)$ is defined for all t in $L^2(\mu)$ by $Q(t) = \|t\|^2 - 2t$. Hence, density estimation is a problem of M-estimation. These problems are classically solved in two steps. First, we choose a "model" S_m that should be close to the parameter s, which means that $\inf_{t \in S_m} \|s - t\|^2$ is "small". Then, we minimize over S_m the empirical version of the integrated contrast, that is, we choose

$$\hat{s}_m \in \arg\min_{t \in S_m} P_n Q(t). \tag{1}$$

This last minimization can be computationaly untractable for general sets S_m, leading to untractable procedures in practice. However, it can be easily solved when S_m is a linear subspace of $L^2(\mu)$ since, for all orthonormal basis $(\psi_\lambda)_{\lambda \in m}$,

$$\hat{s}_m = \sum_{\lambda \in m} (P_n \psi_\lambda)\psi_\lambda. \tag{2}$$

Thus, we will always assume that a model is a linear subspace in $L^2(\mu)$. The risk of the least-squares estimator $\hat{s}_m$ defined in (1) is then decomposed in two terms, called bias and variance, thanks to Pythagoras relation. Let s_m be the orthogonal projection of s onto S_m,

$$\|s - \hat{s}_m\|^2 = \|s - s_m\|^2 + \|s_m - \hat{s}_m\|^2.$$

The statistician should choose a space S_m realizing a trade-off between those terms. S_m must be sufficiently "large" to ensure a small bias $\|s - s_m\|^2$, but not too much, for the variance $\|s_m - \hat{s}_m\|^2$ not to explose. The best trade-off depends on unknown properties of s, since the bias is unknown, and on the behavior of the empirical minimizer $\hat{s}_m$ in the space S_m. Classically, S_m is a parametric space and the dimension d_m of S_m as a linear space is used to give upper bounds on $D_m = n\mathbb{E}\left(\|s_m - \hat{s}_m\|^2\right)$. This approach is validated in regular models under the assumption that the support of s is a known compact, as mentioned in section 3. However, this definition can fail dramatically because there exist simple models (histograms with a small dimension d_m) where D_m is very large, and infinite dimensional models where D_m is easily upper bounded. This issue is extensively discussed in Birgé [7]. Birgé chooses to keep the dimension d_m of S_m as a complexity measure and build new estimators that achieve better risk bounds than the empirical minimizer. His procedures are unfortunatly untractable for the practical user because he can only prove the existence of his estimators. Even his bounds on the risk are only interesting theoretically because they involve constants which are not optimal. We will not take this point of view here and our estimator will always be the empirical minimizer, mainly because it can easily be computed, see (2). We will focus on the quantity D_m/n and introduce a general Assumption (namely Assumption **[V]**) that allows to work

indifferently with D_m/n or with the actual risk $\|s_m - \hat{s}_m\|^2$. We will also provide and study an estimator of D_m/n based on the resampling heuristic.
We insist here on the fact that, unlike classical methods, we will not use in this paper strong extra assumptions on s, like $\|s\|_\infty < \infty$ or assume that s is compactly supported.

1.2 Model selection

Recall that the choice of an optimal model S_m is impossible without strong assumptions on s, for example a precise information on its regularity. However, under less restrictive hypotheses, we can build a countable collection of models $(S_m)_{m\in\mathcal{M}_n}$, growing with the number of observations, such that the best estimator in the associated collection $(\hat{s}_m)_{m\in\mathcal{M}_n}$ realizes an optimal trade-off, see for example Birgé & Massart [8] and Barron, Birgé & Massart [6]. The aim is then to build an estimator $\hat{m}$ such that our final estimator, $\tilde{s} = \hat{s}_{\hat{m}}$ behaves almost as well as any model m_o in the set of oracles

$$\mathcal{M}_n^* = \{m_o \in \mathcal{M}_n,\ \|\hat{s}_{m_o} - s\|^2 = \inf_{m\in\mathcal{M}_n} \|\hat{s}_m - s\|^2\}.$$

This is the problem of model selection. More precisely, we want that $\tilde{s}$ satisfies an oracle inequality defined in general as follows.

Definition: *(Trajectorial oracle inequality) Let $(p_n)_{n\in\mathbb{N}}$ be a summable sequence and let $(C_n)_{n\in\mathbb{N}}$ and $(R_{m,n})_{n\in\mathbb{N}}$ be sequences of positive real numbers. The estimator $\tilde{s} = \hat{s}_{\hat{m}}$ satisfies a trajectorial oracle inequality $TO(C_n, (R_{m,n})_{m\in\mathcal{M}_n}, p_n)$ if*

$$\forall n \in \mathbb{N}^*,\ \mathbb{P}\left(\|\tilde{s} - s\|^2 > C_n \inf_{m\in\mathcal{M}_n} \left\{\|s - \hat{s}_m\|^2 + R_{m,n}\right\}\right) \leq p_n. \tag{3}$$

When $\tilde{s}$ satisfies an oracle inequality, C_n is called the leading constant.

In this paper, we are interested in the problem of optimal model selection defined as follows.

Definition: *(Optimal model selection) We say that $\tilde{s}$ is optimal or that the procedure of selection $(X_1, ..., X_n) \mapsto \hat{m}$ is optimal when $\tilde{s}$ satisfies a trajectorial oracle inequality $TO(1 + r_n, (R_{m,n})_{m\in\mathcal{M}_n}, p_n)$ with $r_n \to 0$ and for all n in $\mathbb{N}^*$ and m in $\mathcal{M}_n$ $R_{m,n} = 0$. In order to simplify the notations, when $\tilde{s}$ is optimal we will say that $\tilde{s}$ satisfies an optimal oracle inequality $OTO(r_n, p_n)$.*

In order to build $\hat{m}$, we remark that, for all m in $\mathcal{M}_n$,

$$\|s - \hat{s}_m\|^2 = \|\hat{s}_m\|^2 - 2P\hat{s}_m + \|s\|^2 = P_nQ(\hat{s}_m) + 2\nu_n(\hat{s}_m) + \|s\|^2, \tag{4}$$

where $\nu_n = P_n - P$ is the centered empirical process. An oracle minimizes $\|s - \hat{s}_m\|^2$ and thus $P_nQ(\hat{s}_m) + 2\nu_n(\hat{s}_m)$. As we want to imitate the oracle, we will design a map $\text{pen} : \mathcal{M}_n \to \mathbb{R}^+$ and choose

$$\hat{m} \in \arg\min_{m\in\mathcal{M}_n} P_nQ(\hat{s}_m) + \text{pen}(m),\ \tilde{s} = \hat{s}_{\hat{m}}. \tag{5}$$

It is clear that the ideal penalty is $\text{pen}_{id}(m) = 2\nu_n(\hat{s}_m)$. For all m in $\mathcal{M}_n$, for all orthonormal basis $(\psi_\lambda)_{\lambda\in m}$, $\hat{s}_m = \sum_{\lambda\in m}(P_n\psi_\lambda)\psi_\lambda$ and $s_m = \sum_{\lambda\in m}(P\psi_\lambda)\psi_\lambda$, thus

$$\nu_n(\hat{s}_m - s_m) = \nu_n\left(\sum_{\lambda\in m}(\nu_n\psi_\lambda)\psi_\lambda\right) = \sum_{\lambda\in m}(\nu_n\psi_\lambda)^2 = \|\hat{s}_m - s_m\|^2.$$

Let us define, for all m in $\mathcal{M}_n$

$$p(m) = \nu_n(\hat{s}_m - s_m) = \|\hat{s}_m - s_m\|^2.$$

From (4), for all m in $\mathcal{M}_n$,

$$\begin{aligned} \|s - \tilde{s}\|^2 &= \|\tilde{s}\|^2 - 2P\tilde{s} + \|s\|^2 = \|\tilde{s}\|^2 - 2P_n\tilde{s} + 2\nu_n\tilde{s} + \|s\|^2 \\ &\leq P_nQ(\hat{s}_m) + \text{pen}(m) + (2\nu_n(\tilde{s}) - \text{pen}(\hat{m})) + \|s\|^2 \\ &= \|s - \hat{s}_m\|^2 + (\text{pen}(m) - 2\nu_n(\hat{s}_m)) + (2\nu_n(\tilde{s}) - \text{pen}(\hat{m})) \end{aligned}$$

Hence, for all m in $\mathcal{M}_n$,

$$\|s - \tilde{s}\|^2 \leq \|s - \hat{s}_m\|^2 + (\text{pen}(m) - 2p(m)) + (2p(\hat{m}) - \text{pen}(\hat{m})) + 2\nu_n(s_{\hat{m}} - s_m). \quad (6)$$

Let us define, for all $c_1, c_2 > 0$, the function

$$f_{c_1,c_2} : \mathbb{R}^+ \to \mathbb{R}^+,\ x \mapsto \begin{cases} \frac{1+c_1x}{1-c_2x} - 1 & \text{if} \quad x < 1/c_2 \\ +\infty & \text{if} \quad x \geq 1/c_2 \end{cases}. \quad (7)$$

It comes from inequality (6) that $\tilde{s}$ satisfies an oracle inequality $OTO(f_{2,2}(\epsilon_n), p_n)$ as soon as, with probability larger than $1 - p_n$

$$\forall m \in \mathcal{M}_n\ \frac{|2p(m) - \text{pen}(m)|}{\|s - \hat{s}_m\|^2} \leq \epsilon_n \text{ and} \quad (8)$$

$$\forall (m, m') \in \mathcal{M}_n^2,\ \frac{2\nu_n(s_{m'} - s_m)}{\|s - \hat{s}_{m'}\|^2 + \|s - \hat{s}_m\|^2} \leq \epsilon_n. \quad (9)$$

Inequality (9) does not depend on our choice of penalty, we will check that it can easily be satisfied in classical collections of models. In order to obtain inequality (8), we use two methods, defined in M-estimation, but studied only on some regression frameworks.

1.2.1 The slope heuristic

The first one is refered as the "slope heuristic". The idea has been introduced by Birgé & Massart [9] in the Gaussian regression framework and developed in a general algorithm by Arlot & Massart [5]. This heuristic states that there exist a sequence $(\Delta_m)_{m \in \mathcal{M}_n}$ and a constant $K_{\min}$ satisfying the following properties,

1. when $\text{pen}(m) < K_{\min}\Delta_m$, then $\Delta_{\hat{m}}$ is too large, typically $\Delta_{\hat{m}} \geq C \max_{m \in \mathcal{M}_n} \Delta_m$,
2. when $\text{pen}(m) \simeq (K_{\min} + \delta)\Delta_m$ for some $\delta > 0$, then $\Delta_{\hat{m}}$ is much smaller,
3. when $\text{pen}(m) \simeq 2K_{\min}\Delta_m$, the selected estimator is optimal.

Thanks to the third point, when Δ_m and $K_{\min}$ are known, this heuristic says that the penalty $\text{pen}(m) = 2K_{\min}\Delta_m$ selects an optimal estimator. When Δ_m only is known, the first and the second point can be used to calibrate $K_{\min}$ in practice, as shown by the following algorithm (see Arlot & Massart [5]):

Slope algorithm
For all $K > 0$, compute the selected model $\hat{m}(K)$ given by (5) with the penalty $\text{pen}(m) = K\Delta_m$ and the associated complexity $\Delta_{\hat{m}(K)}$.
Find the constant $K_{\min}$ such that $\Delta_{\hat{m}(K)}$ is large when $K < K_{\min}$, and "much smaller"

when $K > K_{\min}$.
Take the final $\hat{m} = \hat{m}(2K_{\min})$.

We will justify the slope heuristic in the density estimation framework for $\Delta_m = \mathbb{E}(\|s_m - \hat{s}_m\|^2) = D_m/n$ and $K_{\min} = 1$. In general, D_m is unknown and has to be estimated, we propose a resampling estimator and prove that it can be used without extra assumptions to obtain optimal results.

1.2.2 Resampling penalties

Data-driven penalties have already been used in density estimation in particular cross-validation methods as in Stone [21], Rudemo [20] or Celisse [11]. We are interested here in the resampling penalties introduced by Arlot [4]. Let $(W_1, ..., W_n)$ be a resampling scheme, i.e. a vector of random variables independent of $X, X_1, ..., X_n$ and exchangeable, that is, for all permutations τ of $(1, ..., n)$,

$$(W_1, ..., W_n) \text{ has the same law as } (W_{\tau(1)}, ..., W_{\tau(n)}).$$

Hereafter, we denote by $\bar{W}_n = \sum_{i=1}^n W_i/n$ and by E^W and $\mathcal{L}^W$ respectively the expectation and the law conditionally to the data $X, X_1, ..., X_n$. Let $P_n^W = \sum_{i=1}^n W_i \delta_{X_i}/n$, $\nu_n^W = P_n^W - \bar{W}_n P_n$ be the resampled empirical processes. Arlot's procedure is based on the resampling heuristic of Efron (see Efron [13]), which states that the law of a functional $F(P, P_n)$ is close to its resampled counterpart, that is the conditional law $\mathcal{L}^W(C_W F(\bar{W}_n P_n, P_n^W))$. C_W is a renormalizing constant that depends only on the resampling scheme and on F. Following this heuristic, Arlot defines as a penalty the resampling estimate of the ideal penalty $2D_m/n$, that is

$$\text{pen}(m) = 2C_W \mathbb{E}^W(\nu_n^W(\hat{s}_m^W)), \tag{10}$$

where $\hat{s}_m^W$ minimizes $P_n^W Q(t)$ over S_m. We prove concentration inequalities for pen(m) and deduce that pen(m) provides an optimal procedure.

The paper is organized as follows. In Section 2, we state our main results, we prove the efficiency of the slope algorithm and the resampling penalties.
In Section 3, we compute the rates of convergence in the oracle inequalities using classical collections of models. Section 4 is devoted to a short simulation study where we compare different methods in practice. The proofs are postponed to Section 5. Section 6 is an Appendix where we add some probabilistic material, we prove a concentration inequality for Z^2, where $Z = \sup_{t \in B} \nu_n(t)$ and B is symmetric. We deduce a simple concentration inequality for U-statistics of order 2 that extends a previous result by Houdré & Reynaud-Bouret [16].

2 Main results

Hereafter, we will denote by c, C, K, κ, L, α, with various subscripts some constants that may vary from line to line.

2.1 Concentration of the ideal penalty

Take an orthonormal basis $(\psi_\lambda)_{\lambda \in m}$ of S_m. Easy algebra leads to

$$s_m = \sum_{\lambda \in m} (P\psi_\lambda)\psi_\lambda, \ \hat{s}_m = \sum_{\lambda \in m} (P_n\psi_\lambda)\psi_\lambda, \text{ thus } \|s_m - \hat{s}_m\|^2 = \sum_{\lambda \in m} (\nu_n(\psi_\lambda))^2.$$

$\hat{s}_m$ is an unbiased estimator of s_m and

$$\mathrm{pen}_{id}(m) = 2\nu_n(\hat{s}_m) = 2\nu_n(\hat{s}_m - s_m) + 2\nu_n(s_m) = 2\|s_m - \hat{s}_m\|^2 + 2\nu_n(s_m).$$

For all m, m' in $\mathcal{M}_n$, let

$$p(m) = \|s_m - \hat{s}_m\|^2 = \sum_{\lambda \in m} (\nu_n(\psi_\lambda))^2,\ \delta(m, m') = 2\nu_n(s_m - s_{m'}). \tag{11}$$

From (6), for all m in $\mathcal{M}_n$,

$$\|s - \tilde{s}\|_2^2 \leq \|s - \hat{s}_m\|_2^2 + (\mathrm{pen}(m) - 2p(m)) + (2p(\hat{m}) - \mathrm{pen}(\hat{m})) + \delta(\hat{m}, m). \tag{12}$$

In this section, we are interested in the concentration of $p(m)$ around $\mathbb{E}(p(m)) = D_m/n$. Let us first remark that, for all m in $\mathcal{M}_n$, $p(m)$ is the supremum of the centered empirical process over the ellipsoid $B_m = \{t \in S_m,\ \|t\| \leq 1\}$. From Cauchy-Schwarz inequality, for all real numbers $(b_\lambda)_{\lambda \in m}$,

$$\sum_{\lambda \in m} b_\lambda^2 = \left(\sup_{\sum a_\lambda^2 \leq 1} \sum_{\lambda \in m} a_\lambda b_\lambda \right)^2. \tag{13}$$

We apply this inequality with $b_\lambda = \nu_n(\psi_\lambda)$. We obtain, since the system $(\psi_\lambda)_{\lambda \in m}$ is orthonormal,

$$\sum_{\lambda \in m} (\nu_n(\psi_\lambda))^2 = \sup_{\sum a_\lambda^2 \leq 1} \left(\sum_{\lambda \in m} a_\lambda \nu_n(\psi_\lambda) \right)^2 = \sup_{\sum a_\lambda^2 \leq 1} \left(\nu_n \left(\sum_{\lambda \in m} a_\lambda \psi_\lambda \right) \right)^2 = \sup_{t \in B_m} (\nu_n(t))^2.$$

Hence, $p(m)$ is bounded by a Talagrand's concentration inequality (see Talagrand [22]). This inequality involves $D_m = n\mathbb{E}\left(\|\hat{s}_m - s_m\|^2\right)$ and the constants

$$e_m = \frac{1}{n} \sup_{t \in B_m} \|t\|_\infty^2 \text{ and } v_m^2 = \sup_{t \in B_m} \mathrm{Var}(t(X)). \tag{14}$$

More precisely, the following proposition holds:

Proposition 2.1 *Let $X, X_1, ..., X_n$ be iid random variables with common density s with respect to a probability measure μ. Assume that s belongs to $L^2(\mu)$ and let S_m be a linear subspace in $L^2(\mu)$. Let s_m and $\hat{s}_m$ be respectively the orthogonal projection and the projection estimator of s onto S_m. Let $p(m) = \|s_m - \hat{s}_m\|^2$, $D_m = n\mathbb{E}(p(m))$ and let v_m, e_m be the constants defined in (14). Then, for all $x > 0$,*

$$\mathbb{P}\left(p(m) - \frac{D_m}{n} > \frac{D_m^{3/4}(e_m x^2)^{1/4} + 0.7\sqrt{D_m v_m^2 x} + 0.15 v_m^2 x + e_m x^2}{n} \right) \leq e^{-x/20} \tag{15}$$

$$\mathbb{P}\left(\frac{D_m}{n} - p(m) > \frac{1.8 D_m^{3/4}(e_m x^2)^{1/4} + 1.71\sqrt{D_m v_m^2 x} + 4.06 e_m x^2}{n} \right) \leq 2.8 e^{-x/20} \tag{16}$$

Comments : From (12), for all m in $\mathcal{M}_n$,

$$\begin{aligned} \|s - \tilde{s}\|_2^2 \ \leq\ & \|s - \hat{s}_m\|_2^2 + \left(\mathrm{pen}(m) - 2\frac{D_m}{n} \right) + 2\left(\frac{D_m}{n} - p(m) \right) \\ & + 2\left(p(\hat{m}) - \frac{D_{\hat{m}}}{n} \right) + \left(2\frac{D_{\hat{m}}}{n} - \mathrm{pen}(\hat{m}) \right) + \delta(\hat{m}, m). \end{aligned} \tag{17}$$

It appears from (17) that we can obtain oracle inequalities with a penalty of order $2D_m/n$ if, uniformly over m, m' in $\mathcal{M}_n$,

$$p(m) - \frac{D_m}{n} << \|s - \hat{s}_m\|^2 \text{ and } \delta(m', m) << \|s - \hat{s}_m\|^2 + \|s - \hat{s}_{m'}\|^2.$$

Proposition 2.1 proves that the first part holds with large probability for all m in $\mathcal{M}_n$ such that $e_m \vee v_m^2 << n\mathbb{E}(\|s - \hat{s}_m\|^2)$. Actually, the other part also holds under the same kind of assumption.

2.2 Main assumptions

For all m, m' in $\mathcal{M}_n$, let $D_m = n\mathbb{E}\left(\|s_m - \hat{s}_m\|^2\right)$,

$$\frac{R_m}{n} = \mathbb{E}\left(\|s - \hat{s}_m\|^2\right) = \|s - s_m\|^2 + \frac{D_m}{n},$$

$$v_{m,m'}^2 = \sup_{t \in S_m + S_{m'}, \|t\| \leq 1} \mathrm{Var}(t(X)), e_{m,m'} = \frac{1}{n} \sup_{t \in S_m + S_{m'}, \|t\| \leq 1} \|t\|_\infty^2 .$$

For all $k \in \mathbb{N}$, let $\mathcal{M}_n^k = \{m \in \mathcal{M}_n,\ R_m \in [k, k+1)\}$. For all n in $\mathbb{N}$, for all $k > 0$, $k' > 0$ and $\gamma \geq 0$, let $[k]$ be the integer part of k and let

$$l_{n,\gamma}(k, k') = \ln(1 + \mathrm{Card}(\mathcal{M}_n^{[k]})) + \ln(1 + \mathrm{Card}(\mathcal{M}_n^{[k']})) + \ln((k+1)(k'+1)) + (\ln n)^\gamma \quad (18)$$

Assumption [V]: *There exist $\gamma > 1$ and a sequence $(\epsilon_n)_{n \in \mathbb{N}}$, with $\epsilon_n \to 0$ such that, for all n in $\mathbb{N}$,*

$$\sup_{(k,k') \in (\mathbb{N}^*)^2} \sup_{(m,m') \in \mathcal{M}_n^k \times \mathcal{M}_n^{k'}} \left\{ \left(\left(\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \right)^2 \vee \frac{e_{m,m'}}{R_m \vee R_{m'}} \right) l_{n,\gamma}^2(k, k') \right\} \leq \epsilon_n^4.$$

[BR] *There exist two sequences $(h_n^*)_{n \in \mathbb{N}^*}$ and $(h_n^o)_{n \in \mathbb{N}^*}$ with $(h_n^o \vee h_n^*) \to 0$ as $n \to \infty$ such that, for all n in $\mathbb{N}^*$, for all $m_o \in \arg\min_{m \in \mathcal{M}_n} R_m$ and all $m^* \in \arg\max_{m \in \mathcal{M}_n} D_m$,*

$$\frac{R_{m_o}}{D_{m^*}} \leq h_n^o,\ \frac{n\|s - s_{m^*}\|^2}{D_{m^*}} \leq h_n^*.$$

Comments:

- Assumption **[V]** ensures that the fluctuations of the ideal penalty are uniformly small compared to the risk of the estimator $\hat{s}_m$. Note that for all k, k', $l_{n,\gamma}(k, k') \geq (\ln n)^\gamma$, thus, Assumption **[V]** holds only in typical non parametric situations where $R_n = \inf_{m \in \mathcal{M}_n} R_m \to \infty$ as $n \to \infty$.

- The slope heuristic states that the complexity $\Delta_{\hat{m}}$ of the selected estimator is too large when the penalty term is too small. A minimal assumption for this heuristic to hold with $\Delta_m = D_m$ would be that there exists a sequence $(\theta_n)_{n \in \mathbb{N}^*}$ with $\theta_n \to 0$ as $n \to \infty$ such that, for all n in $\mathbb{N}^*$, for all $m_o \in \arg\min_{m \in \mathcal{M}_n} \mathbb{E}\left(\|s - \hat{s}_m\|^2\right)$ and all $m^* \in \arg\max_{m \in \mathcal{M}_n} \mathbb{E}\left(\|s_m - \hat{s}_m\|^2\right)$,

  $$D_{m_o} \leq \theta_n D_{m^*}.$$

  Assumption **[BR]** is slightly stronger but will always hold in the examples (see Section 3).

In order to have an idea of the rates R_n, ϵ_n, h_n^*, h_n^o and θ_n, let us briefly consider the very simple following example:

Example HR: We assume that s is supported in $[0,1]$ and that $(S_m)_{m\in\mathcal{M}_n}$ is the collection of the regular histograms on $[0,1]$, with $d_m = 1, ..., n$ pieces. We will see in Section 3.2 that $D_m \sim d_m$ asymptotically, hence $D_{m^*} \simeq n$. Moreover, we assume that s is Hölderian and not constant so that there exist positive constants $c_l, c_u, \alpha_l, \alpha_u$ such that, for all m in $\mathcal{M}_n$, see for example Arlot [4],

$$c_l d_m^{-\alpha_l} \leq \|s - s_m\|^2 \leq c_u d_m^{-\alpha_u}.$$

In Section 3.2, we prove that this assumption implies **[V]** with $\epsilon_n \leq C\ln(n)n^{-1/(8\alpha_l+4)}$. Moreover, there exists a constant $C > 0$ such that $R_{m_o} \leq \inf_{m\in\mathcal{M}_n}(c_u n d_m^{-\alpha_u} + d_m) \leq Cn^{-1/(2\alpha_u+1)}$, thus $R_{m_o}/D_m^* \leq Cn^{1/(2\alpha_u+1)-1} = Cn^{-2\alpha_u/(2\alpha_u+1)}$. Since there exists $C > 0$ such that $n\|s - s_{m^*}\|^2/D_{m^*} \leq Cd_{m^*}^{-\alpha_u} = Cn^{-\alpha_u}$, **[BR]** holds with $h_n^o = Cn^{-2\alpha_u/(2\alpha_u+1)}$ and $h_n^* = Cn^{-\alpha_u}$.
Other examples can be found in Birgé & Massart [8], see also Section 3.

2.3 Results on the Slope Heuristic

Let us now turn to the slope heuristic presented in Section 1.2.1.

Theorem 2.2 *(Minimal penalty) Let $\mathcal{M}_n$ be a collection of models satisfying* **[V]** *and* **[BR]** *and let $\epsilon_n^* = \epsilon_n \vee h_n^*$.*
Assume that there exists $0 < \delta_n < 1$ such that $0 \leq pen(m) \leq (1-\delta_n)D_m/n$. Let $\hat{m}, \tilde{s}$ be the random variables defined in (5) and let

$$c_n = \frac{\delta_n - 28\epsilon_n^*}{1 + 16\epsilon_n}.$$

There exists a constant $C > 0$ such that,

$$\mathbb{P}\left(D_{\hat{m}} \geq c_n D_{m^*},\ \|s - \tilde{s}\|^2 \geq \frac{c_n}{5h_n^o}\inf_{m\in\mathcal{M}_n}\|s - \hat{s}_m\|^2\right) \geq 1 - Ce^{-\frac{1}{2}(\ln n)^\gamma}. \tag{19}$$

Comments: Assume that $\text{pen}(m) \leq (1-\delta)D_m/n$, then, inequality (19) proves that an oracle inequality can not be obtained since $c_n/h_n^o \to \infty$. Moreover, $D_{\hat{m}} \geq cD_{m^*}$ is as large as possible. This proves point 1 of the slope heuristic.

Theorem 2.3 *Let $\mathcal{M}_n$ be a collection of models satisfying Assumption* **[V]**. *Assume that there exist $\delta^+ \geq \delta_- > -1$ and $0 \leq p' < 1$ such that, with probability at least $1 - p'$,*

$$2\frac{D_m}{n} + \delta_-\frac{R_m}{n} \leq pen(m) \leq 2\frac{D_m}{n} + \delta^+\frac{R_m}{n}.$$

Let $\hat{m}, \tilde{s}$ be the random variables defined in (5) and let

$$C_n(\delta_-, \delta^+) = \left(\frac{1 + \delta_- - 46\epsilon_n}{1 + \delta^+ + 26\epsilon_n} \vee 0\right)^{-1}.$$

There exists a constant $C > 0$ such that, with probability larger than $1 - p' - Ce^{-\frac{1}{2}(\ln n)^\gamma}$,

$$D_{\hat{m}} \leq C_n(\delta_-, \delta^+)R_{m_o},\ \|s - \tilde{s}\|^2 \leq C_n(\delta_-, \delta^+)\inf_{m\in\mathcal{M}_n}\|s - \hat{s}_m\|^2. \tag{20}$$

Comments :

- Assume that $\text{pen}(m) = KD_m/n$ with $K > 1$, then inequality (20) ensures that $D_{\hat{m}} \leq C_n(K,K)R_{m_o}$. Hence, $D_{\hat{m}}$ jumps from D_{m^*} (Theorem 2.2) to R_{m_o} (20) when $\text{pen}(m)$ is around D_m/n, which is much smaller thanks to Assumption **[BR]**. This proves point 2 of the slope heuristic.

- Point 3 of this heuristic comes from inequality (20) applied with small δ_- and δ^+. The rate of convergence of the leading constant to 1 is then given by the supremum between δ_-, δ^+ and ϵ_n.

- The condition on the penalty has the same form as the one given in Arlot & Massart [5]. It comes from the fact that we do not know D_m/n in many cases, therefore, it has to be estimated. We propose two alternatives to solve this issue. In Section 2.4, we give a resampling estimator of D_m. It can be used for all collection of models satisfying **[V]** and its error of approximation is upper bounded by $\epsilon_n R_m/n$. Thus Theorem 2.3 holds with $(\delta_- \vee \delta^+) \leq C\epsilon_n$. In Section 3.2, we will also see that, in regular models, we can use d_m instead of D_m and the error is upper bounded by CR_m/R_{m_o}, thus Theorem 2.3 holds with $(\delta_- \vee \delta^+) \leq C/R_{m_o} << \epsilon_n$, $p' = 0$. In both cases, we deduce from Theorem 2.3 that the estimator $\tilde{s}$ given by the slope algorithm achieves an optimal oracle inequality $OTO(\kappa\epsilon_n, Ce^{-\frac{1}{2}(\ln n)^\gamma})$. In Example **HR**, for example, we obtain $\epsilon_n = Cn^{-1/(8\alpha_l+4)} \ln n$.

2.4 Resampling penalties

Optimal model selection is possible in density estimation provided that we have a sharp estimation of $D_m = n\mathbb{E}\left(\sup_{t\in B_m}(\nu_n(t))^2\right)$. We propose an estimator of this quantity based on the resampling heuristic. The model selection algorithm that we deduce is the same as the resampling penalization procedure introduced by Arlot [4]. Let F be a fixed functional. Efron's heuristic states that the law $\mathcal{L}(F(\nu_n))$ is close to the conditional law $\mathcal{L}^W(C_W F(\nu_n^W))$, where C_W is a normalizing constant depending only on the resampling scheme and the functional F. Let $P_n^W = \sum_{i=1}^n W_i\delta_{X_i}/n$ and $\nu_n^W = P_n^W - \bar{W}_n P_n$. The resampling estimator of D_m is $D_m^W = nC_W^2\mathbb{E}^W\left(\sup_{t\in B_m}(\nu_n^W(t))^2\right)$ and the resampling penalty associated is $\text{pen}(m) = 2D_m^W/n$. Actually, the following result describes the concentration of D_m^W around its mean D_m and around $np(m)$.

Proposition 2.4 *Let $(W_1, ..., W_n)$ be a resampling scheme, let S_m be a linear space, $B_m = \{t \in S_m,\ \|t\| \leq 1\}$, $p(m) = \sup_{t\in B_m}(\nu_n(t))^2$, $D_m = n\mathbb{E}\left(p(m)\right)$ and let D_m^W be the resampling estimator of D_m based on $(W_1, ..., W_n)$, that is $D_m^W = nC_W^2\mathbb{E}^W\left(\sup_{t\in B_m}(\nu_n^W(t))^2\right)$, where $v_W^2 = Var(W_1 - \bar{W}_n)$ and $C_W^2 = (v_W^2)^{-1}$.*
Then, for all m in $\mathcal{M}_n$, $\mathbb{E}(D_m^W) = D_m$. Moreover, let e_m, v_m be the quantities defined in (14). For all $x > 0$, on an event of probability larger than $1 - 7.8e^{-x}$,

$$\begin{aligned} D_m^W - D_m &\leq \sqrt{8e_m D_m x} + e_m\left(\frac{4x}{3} + \frac{(40.3x)^2}{n-1}\right) \\ &\quad + \frac{9D_m^{3/4}(e_m x^2)^{1/4} + 7.61\sqrt{v_m^2 D_m x}}{n-1}. \qquad (21) \\ D_m^W - D_m &\geq -\sqrt{8e_m D_m x} - e_m\left(\frac{4x}{3} + \frac{(19.1x)^2}{n-1}\right) \\ &\quad - \frac{5.31D_m^{3/4}(e_m x^2)^{1/4} + 3\sqrt{v_m^2 D_m x} + 3v_m^2 x}{n-1}. \qquad (22) \end{aligned}$$

For all $x > 0$,

$$\mathbb{P}\left(p(m) - \frac{D_m^W}{n} > \frac{5.31 D_m^{3/4}(e_m x^2)^{1/4} + 3\sqrt{v_m^2 D_m x} + 3v_m^2 x + e_m(19.1x)^2}{n-1}\right) \leq 2e^{-x} \tag{23}$$

$$\mathbb{P}\left(\frac{D_m^W}{n} - p(m) \leq \frac{9 D_m^{3/4}(e_m x^2)^{1/4} + 7.61\sqrt{v_m^2 D_m x} + e_m(40.3x)^2}{n-1}\right) \leq 3.8e^{-x}. \tag{24}$$

Remark
The concentration of the resampling estimator involves the same quantities as the concentration of $p(m)$, thus, it can be used to estimate the ideal penalty in the slope heuristic's algorithm presented in the previous section without extra assumptions on the collection $\mathcal{M}_n$. Proposition 2.4 and Theorem 2.3 prove that this resampling penalty leads to an efficient model selection procedure. However, we do not need to use the slope heuristic in our framework to obtain an optimal model selection procedure as shown by the following theorem.

Theorem 2.5 *Let $X_1, ..., X_n$ be i.i.d random variables with common density s. Let $\mathcal{M}_n$ be a collection of models satisfying Assumption* **[V]**. *Let $W_1, ..., W_n$ be a resampling scheme, let $\bar{W}_n = \sum_{i=1}^n W_i/n$, $v_W^2 = Var(W_1 - \bar{W}_n)$ and $C_W = 2(v_W^2)^{-1}$. Let $\tilde{s}$ be the penalized least-squares estimator defined in (5) with*

$$pen(m) = C_W \mathbb{E}^W\left(\sup_{t \in B_m} (\nu_n^W(t))^2\right).$$

Then, there exists a constant $C > 0$ such that

$$\mathbb{P}\left(\|s - \tilde{s}\|^2 \leq (1 + 100\epsilon_n) \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2\right) \geq 1 - Ce^{-\frac{1}{2}(\ln n)^\gamma}. \tag{25}$$

Comments : The main advantage of this results is that the penalty term is always totally computable. Unlike the penalties derived from the slope heuristic, it does not depend on an arbitrary choice of a constant $K_{\min}$ made by the observer, that may be hard to detect in practice (see the paper of Alot & Massart [5] for an extensive discussion on this important issue). However, C_W is only optimal asymptotically. It is sometimes useful to overpenalize a little in order to improve the non-asymptotic performances of our procedures (see Massart [19]) and the slope heuristic can be used to do it in an optimal way (see our short simulation study in Section 4).

2.5 A remarks on the "regularization phenomenon"

The regularization of the bootstrap phenomenon (see Arlot [3, 4] and the references therein) states that the resampling estimator $C_W \mathbb{E}^W(F(\nu_n^W))$ of a functional $F(\nu_n)$ concentrates around its mean better than $F(\nu_n)$. This phenomenon can be justified with our previous results for our functional F. Recall that we have proven in Proposition 2.1 that, for all $x > 0$, with probability larger than $1 - 3.8e^{-x/20}$,

$$\left|p(m) - \frac{D_m}{n}\right| \leq \frac{1.8 D_m^{3/4}(e_m x^2)^{1/4} + 1.5\sqrt{D_m v_m^2 x} + 0.2 v_m^2 x + 4.1 e_m x^2}{n}.$$

In Example **HR**, we have the following upper bounds

$$D_m \leq d_m,\ e_m \leq \frac{d_m}{n},\ v_m^2 \leq c\|s\|\sqrt{d_m}.$$

Thus, there exists a constant C such that, for all $x > 0$,

$$\mathbb{P}\left(|np(m) - D_m| > Cd_m\left(\sqrt{\frac{x}{\sqrt{n}}} + \left(\frac{x}{\sqrt{n}}\right)^2\right)\right) \leq 3.8e^{-x/20}. \tag{26}$$

On the other hand, it comes from Inequalities (21) and (22), that, for all $x > 0$, on an event of probability larger than $1 - 7.8e^{-x/20}$,

$$\begin{aligned}\left|D_m^W - D_m\right| &\leq \sqrt{0.4e_m D_m x} + e_m\left(\frac{x}{15} + \frac{4.1x^2}{n-1}\right) \\ &\quad + \frac{1.8D_m^{3/4}(e_m x^2)^{1/4} + 1.45\sqrt{v_m^2 D_m x} + 0.2v_m^2 x}{n-1}.\end{aligned}$$

Thus, there exists a constant C such that, for all $x > 0$,

$$\mathbb{P}\left(\left|D_m^W - D_m\right| > Cd_m\left(\sqrt{\frac{x}{n}} + \left(\frac{x}{n}\right)^2\right)\right) \leq 7.8e^{-x/20}.$$

The concentration of D_m^W is then much better than the one of $np(m)$. This implies that D_m^W is an estimator of D_m rather than an estimator of $np(m)$. Thus, the resampling penalty can be used when D_m/n is a good penalty for example, under **[V]**. When D_m/n is known to underpenalize (see the examples in Barron, Birgé & Massart [6]), there is no chance that D_m^W/n can work.

3 Rates of convergence for classical examples

The aim of this section is to show that **[V]** can be derived from a more classical hypothesis in two classical collections of models: the histograms and Fourier spaces. We derive the rates ϵ_n under this new hypothesis.

3.1 Assumption on the risk of the oracle

As mentioned in Section 2.2, Assumption **[V]** can only hold if there exists $\gamma > 1$ such that $R_n(\ln n)^{-\gamma} \to \infty$ as $n \to \infty$, where $R_n = \inf_{m \in \mathcal{M}_n} R_m$. In our example, we will make the following Assumption that ensures that this condition is always satisfied.

[BR] *(Bounds on the Risk) There exist constants $C_u > 0$, $\alpha_u > 0$, $\gamma > 1$, and a sequence $(\theta_n)_{n\in\mathbb{N}}$ with $\theta_n \to \infty$ as $n \to \infty$ such that, for all n in $\mathbb{N}^*$, for all m in $\mathcal{M}_n$*

$$\theta_n^2(\ln n)^{2\gamma} \leq R_n \leq R_m \leq C_u n^{\alpha_u}.$$

Comments: Assumption **[BR]** holds with $\theta_n = Cn^\alpha$ for the collection of regular histograms of example **HR**, provided that s is an Hölderian, non constant and compactly supported function (see for example Arlot [3]). It is also a classical result of minimax theory that there exist functions in Sobolev spaces satisfying this kind of Assumption when $\mathcal{M}_n$ is the collection of Fourier spaces that we will introduce below.

We want to check that these collections satisfy Assumption **[V]**, i.e. that there exists $\gamma > 1$ such that

$$\sup_{(k,k')\in(\mathbb{N}^*)^2} \sup_{(m,m')\in\mathcal{M}_n^k\times\mathcal{M}_n^{k'}} \left\{\left(\left(\frac{v_{m,m'}^2}{R_m \vee R_{m'}}\right)^2 \vee \frac{e_{m,m'}}{R_m \vee R_{m'}}\right) l_{n,\gamma}^2(k,k')\right\} \leq \epsilon_n^4.$$

For all $m \in \mathcal{M}_n$, $R_m \leq C_u n^{\alpha_u}$, thus for all $k > C_u n^{\alpha_u}$, $\mathrm{Card}(\mathcal{M}_n^k) = 0$. In particular, we can assume in the previous supremum that $k \leq C_u n^{\alpha_u}$ and $k' \leq C_u n^{\alpha_u}$. Hence, there exists a constant $\kappa > 0$ such that $\ln[(1+k)(1+k')] \leq \kappa \ln n$. We also add the following assumption that ensures that there exists a constant $\kappa > 0$ such that, for all $k \in \mathbb{N}$, $\ln(1 + Card(\mathcal{M}_n^k)) \leq \kappa \ln n$.

[PC] *(Polynomial collection) There exist constants $c_{\mathcal{M}} \geq 0$, $\alpha_{\mathcal{M}} \geq 0$, such that, for all n in $\mathbb{N}$,*

$$\mathrm{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}.$$

Under Assumptions **[BR]** and **[PC]**, there exists a constant $\kappa > 0$ such that, for all $\gamma > 1$ and $n \geq 3$,

$$\sup_{(k,k') \in (\mathbb{N}^*)^2} \sup_{(m,m') \in \mathcal{M}_n^k \times \mathcal{M}_n^{k'}} \left\{ \left(\left(\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \right)^2 \vee \frac{e_{m,m'}}{R_m \vee R_{m'}} \right) l_{n,\gamma}^2(k,k') \right\}$$
$$\leq \sup_{(m,m') \in (\mathcal{M}_n)^2} \left\{ \left(\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \right)^2 \vee \frac{e_{m,m'}}{R_m \vee R_{m'}} \right\} \kappa (\ln n)^{2\gamma}.$$

3.2 The histogram case

Let $(\mathbb{X}, \mathcal{X})$ be a measurable space. Let $(P_m)_{m \in \mathcal{M}_n}$ be a collection of measurable partitions $P_m = (I_\lambda)_{\lambda \in m}$ of subsets of $\mathbb{X}$ such that, for all $m \in \mathcal{M}_n$, for all $\lambda \in m$, $0 < \mu(I_\lambda) < \infty$. Let m in $\mathcal{M}_n$, the set S_m of histograms associated to P_m is the set of functions which are constant on each I_λ, $\lambda \in m$. S_m is a linear space. Setting, for all $\lambda \in m$, $\psi_\lambda = (\sqrt{\mu(I_\lambda)})^{-1} 1_{I_\lambda}$, the functions $(\psi_\lambda)_{\lambda \in m}$ form an orthonormal basis of S_m.
Let us recall that, for all m in $\mathcal{M}_n$,

$$D_m = \sum_{\lambda \in m} \mathrm{Var}(\psi_\lambda(X)) = \sum_{\lambda \in m} P(\psi_\lambda^2) - (P\psi_\lambda)^2 = \sum_{\lambda \in m} \frac{P(X \in I_\lambda)}{\mu(I_\lambda)} - \|s_m\|^2. \tag{27}$$

Moreover, from Cauchy-Schwarz inequality, for all x in $\mathbb{X}$, for all m, m' in $\mathcal{M}_n$

$$\sup_{t \in B_{m,m'}} t^2(x) \leq \sum_{\lambda \in m \cup m'} \psi_\lambda^2(x), \text{ thus } e_{m,m'} = \frac{1}{n} \sup_{\lambda \in m \cup m'} \frac{1}{\mu(I_\lambda)}. \tag{28}$$

Finally, it is easy to check that, for all m ,m' in $\mathcal{M}_n$

$$v_{m,m'}^2 = \sup_{\lambda \in m \cup m'} \mathrm{Var}(\psi_\lambda(X)) = \sup_{\lambda \in m \cup m'} \frac{P(X \in I_\lambda)(1 - P(X \in I_\lambda))}{\mu(I_\lambda)}. \tag{29}$$

We will consider two particular types of histograms.
Example 1 [Reg] : μ-regular histograms.
For all m in $\mathcal{M}_n$, P_m is a partition of $\mathbb{X}$ and there exist a family $(d_m)_{m \in \mathcal{M}_n}$ bounded by n and two constants c_{rh}, C_{rh} such that, for all m in $\mathcal{M}_n$, for all $\lambda \in \mathcal{M}_n$,

$$\frac{c_{rh}}{d_m} \leq \mu(I_\lambda) \leq \frac{C_{rh}}{d_m}.$$

The typical example here is the collection described in Example **HR**.

Example 2 [Ada]: Adapted histograms.

There exist positive constants c_r, C_{ah} such that, for all m in $\mathcal{M}_n$, for all $\lambda \in \mathcal{M}_n$, $\mu(I_\lambda) \geq c_r n^{-1}$ and

$$\frac{P(X \in I_\lambda)}{\mu(I_\lambda)} \leq C_{ah}.$$

[Ada] is typically satisfied when s is bounded on $\mathbb{X}$. Remark that the models satisfying **[Ada]** have finite dimension $d_m \leq Cn$ since

$$1 \geq \sum_{\lambda \in m} P(X \in I_\lambda) \geq C_{ah} \sum_{\lambda \in m} \mu(I_\lambda) \geq C_{ah} c_r d_m n^{-1}.$$

The example [Reg].
It comes from equations (27, 28, 29) and Assumption **[Reg]** that

$$C_{rh}^{-1} d_m - \|s_m\|^2 \leq D_m \leq c_{rh}^{-1} d_m - \|s_m\|^2.$$

$$e_{m,m'} \leq c_{rh}^{-1} \frac{d_m \vee d_{m'}}{n},\ v_{m,m'}^2 \leq \sup_{t \in B_{m,m'}} \|t\|_\infty \|t\| \|s\| \leq c_{rh}^{-1/2} \|s\| \sqrt{d_m \vee d_{m'}}.$$

Thus

$$\frac{e_{m,m'}}{R_m \vee R_{m'}} \leq C_{rh} c_{rh}^{-1} \frac{(R_m \vee R_{m'}) + \|s\|^2}{n(R_m \vee R_{m'})} \leq C n^{-1}.$$

If $D_m \vee D_{m'} \leq \theta_n^2 (\ln n)^{2\gamma}$,

$$\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \leq \sqrt{C_{rh} c_{rh}^{-1}} \frac{\sqrt{(D_m \vee D_{m'}) + \|s\|^2}}{R_{m_o}} \leq \frac{C}{\theta_n (\ln n)^\gamma}.$$

If $D_m \vee D_{m'} \geq \theta_n^2 (\ln n)^{2\gamma}$,

$$\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \leq \sqrt{C_{rh} c_{rh}^{-1}} \frac{\sqrt{(D_m \vee D_{m'}) + \|s\|^2}}{D_m \vee D_{m'}} \leq \frac{C}{\theta_n (\ln n)^\gamma}.$$

There exists $\kappa > 0$ such that $\theta_n^2 (\ln n)^{2\gamma} \leq \kappa n$ since for all m in $\mathcal{M}_n$, $R_m \leq n\|s - s_m\|^2 + c_{rh}^{-1} d_m \leq (\|s\|^2 + c_{rh}^{-1}) n$. Hence Assumption **[V]** holds with γ given in Assumption **[BR]** and $\epsilon_n = C\theta_n^{-1/2}$.

The example [Ada].
It comes from inequalities (28), (29) and Assumption **[Ada]** that, for all m and m' in $\mathcal{M}_n$

$$e_{m,m'} \leq c_r^{-1} \text{ and } v_{m,m'}^2 \leq C_{ah}.$$

Thus, there exists a constant $\kappa > 0$ such that, for all m an m' in $\mathcal{M}_n$,

$$\sup_{(m,m') \in (\mathcal{M}_n)^2} \left\{ \left(\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \right)^2 \vee \frac{e_{m,m'}}{R_m \vee R_{m'}} \right\} \leq \frac{\kappa}{\theta_n^2 (\ln n)^{2\gamma}}.$$

Therefore Assumption **[V]** holds also with γ given in Assumption **[BR]** and $\epsilon_n = \kappa \theta_n^{-1/2}$.

3.3 Fourier spaces

In this section, we assume that s is supported in $[0,1]$. We introduce the classical Fourier basis. Let $\psi_0 : [0,1] \to \mathbb{R},\ x \mapsto 1$ and, for all $k \in \mathbb{N}^*$, we define the functions

$$\psi_{1,k} : [0,1] \to \mathbb{R},\ x \mapsto \sqrt{2}\cos(2\pi kx),\ \psi_{2,k} : [0,1] \to \mathbb{R},\ x \mapsto \sqrt{2}\sin(2\pi kx).$$

For all j in $\mathbb{N}^*$, let

$$m_j = \{0\} \cup \{(i,k),\ i = 1,2,\ k = 1,...,j\} \text{ and } \mathcal{M}_n = \{m_j, j = 1,...,n\}.$$

For all m in $\mathcal{M}_n$, let S_m be the space spanned by the family $(\psi_\lambda)_{\lambda\in m}$. $(\psi_\lambda)_{\lambda\in m}$ is an orthonormal basis of S_m and for all j in $1,...,n$, $d_{m_j} = 2j+1$.
Let j in $1,...n$, for all x in $[0,1]$,

$$\sum_{\lambda\in m_j} \psi_\lambda^2(x) = 1 + 2\sum_{k=1}^{j} \cos^2(2\pi kx) + \sin^2(2\pi kx) = 1 + 2j = d_{m_j}.$$

Hence, for all m in $\mathcal{M}_n$,

$$D_m = P\left(\sum_{\lambda\in m_j} \psi_\lambda^2\right) - \|s_m\|^2 = d_m - \|s_m\|^2. \tag{30}$$

It is also clear that, for all m, m' in $\mathcal{M}_n$,

$$e_{m,m'} = \frac{d_m \vee d_{m'}}{n},\ v_{m,m'}^2 \le \|s\|\sqrt{d_m \vee d_{m'}}. \tag{31}$$

The collection of Fourier spaces of dimension $d_m \le n$ satisfies Assumption **[PC]**, and the quantities D_m $e_{m,m'}$ and $v_{m,m'}^2$ satisfy the same inequalities as in the collection **[Reg]**, therefore, **[V]** comes also in this collection from **[BR]**. We have obtained the following corollary of Theorem 2.5.

Corollary 3.1 *Let $\mathcal{M}_n$ be either a collection of histograms satisfying Assumptions* **[PC]-[Reg]** *or* **[PC]-[Ada]** *or the collection of Fourier spaces of dimension $d_m \le n$. Assume that s satisfies Assumption* **[BR]** *for some $\gamma > 1$ and $\theta_n \to \infty$. Then, there exist constants $\kappa > 0$ and $C > 0$ such that the estimator $\tilde{s}$ selected by a resampling penalty satisfies*

$$\mathbb{P}\left(\|s - \tilde{s}\|^2 \le (1 + \kappa\theta_n^{-1/2}) \inf_{m\in\mathcal{M}_n} \|s - \hat{s}_m\|^2\right) \ge 1 - Ce^{-\frac{1}{2}(\ln n)^\gamma}.$$

Comment: Assumption **[BR]** is hard to check in practice. We mentioned that it holds in Example **HR** provided that s is Hölderian, non constant and compactly supported (see Arlot [4]). It is also classical to build functions satisfying **[BR]** with the Fourier spaces in order to prove that the oracle reaches the minimax rate of convergence over some Sobolev balls, see for example Birgé & Massart [8], Barron, Birgé & Massart [6] or Massart [19]. In these cases, there exist $c > 0$, $\alpha > 0$ such that $\theta_n \ge cn^\alpha$. In more general situations, we can use the same trick as Arlot [4] and use our main theorem only for the models with dimension $d_m \ge (\ln n)^{4+2\gamma}$, they satisfy **[BR]** with $\theta_n = (\ln n)^2$, at least when n is sufficiently large, because

$$\|s\|^2 + R_m \ge \|s\|^2 + D_m \ge cd_m \ge c(\ln n)^4(\ln n)^{2\gamma}.$$

With our concentration inequalities, we can control easily the risk of the models with dimension $d_m \le (\ln n)^{4+2\gamma}$ by $\kappa(\ln n)^{3+5\gamma/2}$ with probability larger than $1 - Ce^{-\frac{1}{2}(\ln n)^\gamma}$ and we can then deduce the following corollary.

Corollary 3.2 *Let $\mathcal{M}_n$ be either a collection of histograms satisfying Assumptions* **[PC]-[Reg]** *or* **[PC]-[Ada]** *or the collection of Fourier spaces of dimension $d_m \leq n$. There exist constants $\kappa > 0$, $\eta > 3 + 5\gamma/2$ and $C > 0$ such that the estimator $\tilde{s}$ selected by a resampling penalty satisfies*

$$\mathbb{P}\left(\|s - \tilde{s}\|^2 \leq (1 + \kappa(\ln n)^{-1})\left(\inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2 + \frac{(\ln n)^\eta}{n}\right)\right) \geq 1 - Ce^{-\frac{1}{2}(\ln n)^\gamma}.$$

4 Simulation study

We propose in this section to show the practical performances of the slope algorithm and the resampling penalties on two examples. We estimate the density

$$s(x) = \frac{3}{4}x^{-1/4}1_{[0,1]}(x)$$

and we compare the three following methods.

1. The first one is the slope heuristic applied with the linear dimension d_m of the models. We observe two main behaviors of $d_{\hat{m}(K)}$ with respect to K. Most of the times, we only observe one jump, as in Figure 1, and we find $K_{\min}$ easily.

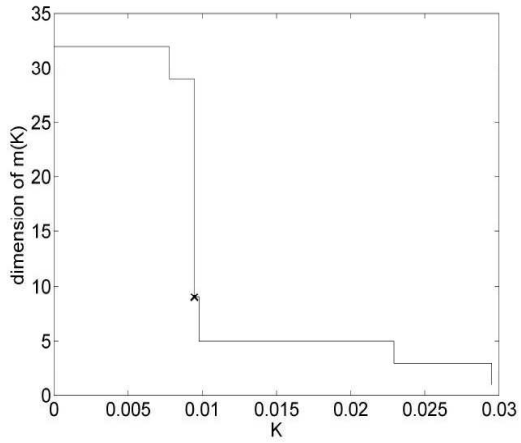

Figure 1: Classical behavior of $K \mapsto d_{\hat{m}(K)}$

We also observe more difficult situations as the one of Figure 2 below, where we can see several jumps. In these cases, as prescribed in the regression framework by Arlot & Massart [5], we choose the constant $K_{\min}$ realizing the maximal jump of $d_{\hat{m}(K)}$. Arlot & Massart [5] also proposed to select $K_{\min}$ as the minimal K such that $d_{\hat{m}(K)} \leq d_{m^*}(\ln n)^{-1}$, but they obtained worse performances of the selected estimator in their simulations.
We justify this method only for collection of models where $d_m \simeq KD_m$ for some constant K. We will see that it gives really good performances when this condition is satisfied.

2. The second method is the resampling based penalization algorithm of Theorem 2.5. Note here that all the resampling penalties D_m^W/n can be easily computed, without any Monte Carlo approximations. Actually, for all resampling scheme,

$$\frac{D_m^W}{n} = \frac{1}{n}\sum_{\lambda \in m}\left(P_n\psi_\lambda^2 - \frac{1}{n(n-1)}\sum_{i \neq j=1}^{n}\psi_\lambda(X_i)\psi_\lambda(X_j)\right).$$

Resampling penalties give always good approximations of D_m. However, in non asymptotic situations, it may be usefull to overpenalize a little bit in order to improve the leading constants in the oracle inequality (in Theorem 2.3, imagine that $46\epsilon_n$ is very close to 1).

3. In a third method, we propose therefore to use the slope algorithm applied with a complexity D_m^W. By this way, we hope to overpenalize a little bit the resampling penalty when it is necessary.

4.1 Example 1: regular case

In the first example, we consider the collection of regular histograms described in example **HR** and we observe $n = 100$ data. In this example, we saw that $D_m^W \simeq D_m \simeq d_m$. We can actually verify in Figure 2 that these quantities almost coincide for the selected model.

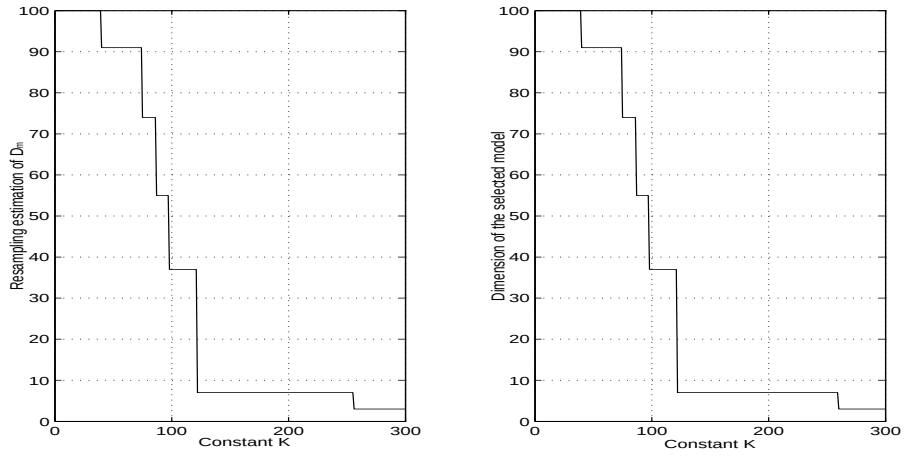

Figure 2: Comparison of d_m and D_m^W on the selected model

We compute $N = 1000$ times the oracle constant $c = \|s - \tilde{s}\|^2/(\inf_{m\in\mathcal{M}_n} \|s - \hat{s}_m\|^2)$ for the 3 methods. We put in the following array the mean, the median and the 0.95-quantile, $q_{0.95}$ of these quantities.

method	mean of the N constants c	median	$q_{0.95}$
slope + $\mathrm{d_m}$	3.56	2.30	10.07
resampling	4.43	2.52	15.47
resampling + slope	3.57	2.21	10.86

We observe that the slope algorithm allows to improve the resampling penalty in practice. This may be due to a little overpenalization even if it is not a straightforward consequence of our theoretical results. Note that, as $d_m \simeq D_m^W$, the slope algorithm leads to the same results when applied with d_m or with D_m^W. Although we have an explicite formula to compute the resampling penalties, the computation time is much longer if we use D_m^W. Therefore, we clearly recommand to use the slope algorithm with d_m for regular collections of model, as regular histograms or Fourier spaces described in Section 3.3.

4.2 Example 2: a more complicated collection

In the next example, we want to show that the linear dimension shall not be used in general. Let us consider a slightly more complicated collection. Let k, J_1, J_2, n be four

non null integers satisfying $k \leq n$, $J_1 \leq k$, $J_2 \leq n-k$. We denote by $S_{k,J_1,J_2,n}$ the linear space of histograms on the following partition.

$$\left\{ \left[l\frac{k}{J_1 n}, (l+1)\frac{k}{J_1 n} \right[,\ l = 0, ..., J_1 - 1 \right\}$$
$$\cup \left\{ \left[\frac{k}{n} + l\frac{1-k/n}{J_2}, \frac{k}{n} + (l+1)\frac{1-k/n}{J_2} \right[,\ l = 0, ...J_2 - 1 \right\}.$$

Let $n \in \mathbb{N}^*$ and let $\mathcal{M}_n = \{(k, J_1, J_2) \in (\mathbb{N}^*)^3;\ k \leq n,\ J_1 \leq k,\ J_2 \leq n-k\}$. It is clear that $\text{Card}(\mathcal{M}_n) \leq n^3$. The oracle of this collection is better than the previous one since the regular histograms belongs to $(S_{m,n})_{m\in\mathcal{M}_n}$. It is easy to check that the dimension of $S_{k,J_1,J_2,n}$ is equal to $J_1 + J_2$ and that $D_{k,J_1,J_2,n}$ is equal to $(nJ_1/k)F(k/n) + (nJ_2/(n-k))(1-F(k/n)) - \|s_{k,J_1,J_2,n}\|^2/n$, where F is the distribution function of the observations. Hence, there is no constant K_o such that $K_o d_{k,J_1,J_2,n} \simeq D_{k,J_1,J_2,n}$ as in the previous example. Figure 3 let us see this fact on the selected model.

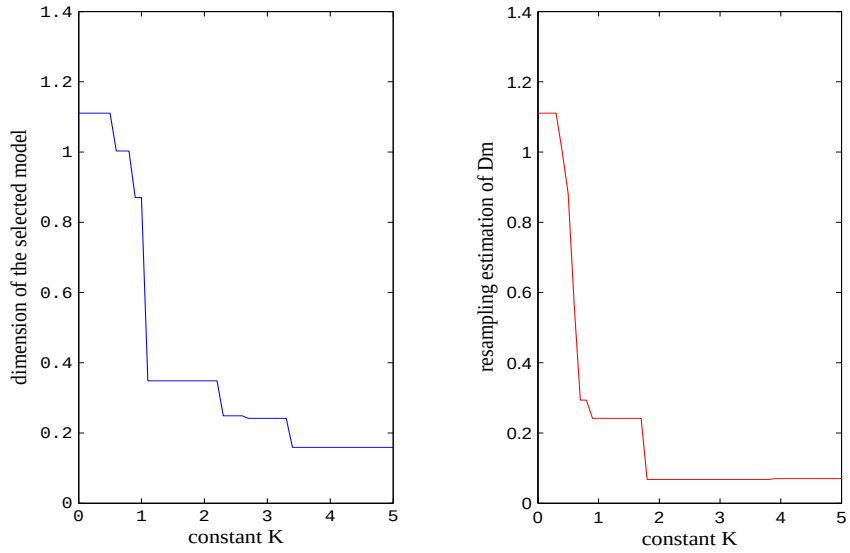

Figure 3: Comparison of d_m and D_m^W on the selected model

We also compute $N = 1000$ times the oracle constant $c = \|s - \tilde{s}\|^2/(\inf_{m\in\mathcal{M}_n} \|s - \hat{s}_m\|^2)$ for the 3 methods, taking $n = 100$ observations each time. The results are summarized in following array.

method	mean of the N constants c	median	$q_{0.95}$
slope + d_m	8.30	7.01	19.73
resampling	6.11	5.08	13.52
resampling + slope	5.33	4.04	12.92

The slope heuristic gives bad results when applied with d_m. This is due to the fact that d_m is not proportional to D_m here. The resampling based penalty $2D_m^W/n$ is much better and, as in the regular case, it is well improved by the slope algorithm. Therefore, for general collections of models where we do not know an optimal shape of the ideal penalty, we recommand to apply the slope algorithm with a complexity equal to D_m^W.

5 Proofs

5.1 Proof of Proposition 2.1

It is a straightforward application of Corollary 6.6 in the appendix.

5.2 Technical lemmas

Before giving the proofs of the main theorems, we state and prove some important technical lemmas that we will use repeatedly all along the proofs. Let us recall here the main notations. For all m, m' in $\mathcal{M}_n$,

$$p(m) = \|s_m - \hat{s}_m\|^2,\ D_m = n\mathbb{E}(p(m)) = n\mathbb{E}\left(\|\hat{s}_m - s_m\|^2\right)$$

$$R_m = n\mathbb{E}\left(\|s - \hat{s}_m\|^2\right) = n\|s - s_m\|^2 + D_m,\ \delta(m, m') = \nu_n(s_m - s_{m'}).$$

For all $n \in \mathbb{N}^*,\ k > 0,\ k' > 0,\ \gamma > 0,$, let $[k]$ be the integer part of k and let

$$l_{n,\gamma}(k, k') = \ln((1 + \mathrm{Card}(\mathcal{M}_n^{[k]}))(1 + \mathrm{Card}(\mathcal{M}_n^{[k']}))) + \ln((1+k)(1+k')) + (\ln n)^\gamma.$$

Recall that Assumption **[V]** implies that, for all m, m' in $\mathcal{M}_n$,

$$\begin{aligned} v_{m,m'}^2 l_{n,\gamma}(R_m, R_{m'}) &\leq \epsilon_n^2(R_m \vee R_{m'}), \\ e_{m,m'}(l_{n,\gamma}(R_m, R_{m'}))^2 &\leq \epsilon_n^4(R_m \vee R_{m'}). \end{aligned} \tag{32}$$

Let us prove a simple result

Lemma 5.1 *For all $K > 1$,*

$$\Sigma(K) = \sum_{k \in \mathbb{N}} \sum_{m \in \mathcal{M}_n^k} e^{-K[\ln(1+\mathrm{Card}(\mathcal{M}_n^k)) + \ln(1+k)]} < \infty. \tag{33}$$

For all m in $\mathcal{M}_n$, let $l_m = l_{n,\gamma}(R_m, R_m)$, then, for all $K > 1/\sqrt{2}$,

$$\sum_{m \in \mathcal{M}_n} e^{-K^2 l_m} = \Sigma(2K^2) e^{-K^2 (\ln n)^\gamma}. \tag{34}$$

For all m, m' in $\mathcal{M}_n$, let $l_{m,m'} = l_{n,\gamma}(R_m, R_{m'})$, then, for all $K > 1$,

$$\sum_{(m,m') \in (\mathcal{M}_n)^2} e^{-K^2 l_{m,m'}} = (\Sigma(K^2))^2 e^{-K^2 (\ln n)^\gamma}. \tag{35}$$

Proof :

Inequality (33) comes from the fact that, when $K > 1$,

$$\forall k \in \mathbb{N},\ \sum_{m \in \mathcal{M}_n^k} e^{-K[\ln(1+\mathrm{Card}(\mathcal{M}_n^k))]} \leq 1,\ \text{and}\ \sum_{k \in \mathbb{N}^*} e^{-K \ln k} < \infty.$$

For all integer k such that $\mathcal{M}_n^k \neq \emptyset$, for all m in $\mathcal{M}_n^k$, $l_m \geq 2[\ln(1 + \mathrm{Card}(\mathcal{M}_n^k)) + \ln(1 + k)] + (\ln n)^\gamma$, thus, for all $K > 1/\sqrt{2}$, it comes from (33) that

$$\sum_{m \in \mathcal{M}_n} e^{-K^2 l_m} \leq e^{-K^2 (\ln n)^\gamma} \sum_{k \in \mathbb{N}} \sum_{m \in \mathcal{M}_n^k} e^{-2K^2[\ln(1+\mathrm{Card}(\mathcal{M}_n^k)) + \ln(1+k)]} \leq \Sigma(2K^2) e^{-K^2 (\ln n)^\gamma}.$$

Finally, for all integers (k, k') such that $\mathcal{M}_n^k \times \mathcal{M}_n^{k'} \neq \emptyset$,

$$l_{m,m'} \geq \ln(1 + \mathrm{Card}(\mathcal{M}_n^k)) + \ln(1+k) + \ln(1 + \mathrm{Card}(\mathcal{M}_n^{k'})) + \ln(1+k') + (\ln n)^\gamma.$$

Thus, from (33),

$$\sum_{(m,m')\in(\mathcal{M}_n^2)} e^{-K^2 l_{m,m'}} = \left(\sum_{k\in\mathbb{N}} \sum_{m\in\mathcal{M}_n^k} e^{-K^2[\ln(1+\mathrm{Card}(\mathcal{M}_n^k))+\ln(1+k)]} \right)^2 e^{-K^2(\ln n)^\gamma}.$$

Lemma 5.2 *Let $\mathcal{M}_n$ be a collection of models satisfying Assumption* **[V]**. *We consider the following events.*

$$\begin{aligned}
\Omega_\delta &= \left\{ \forall (m,m') \in \mathcal{M}_n^2,\ \delta(m,m') \leq 6\epsilon_n \frac{R_m \vee R_{m'}}{n} \right\} \\
\Omega_p &= \bigcap_{m\in\mathcal{M}_n} \left\{ \left\{ p(m) - \frac{D_m}{n} \leq 10\epsilon_n \frac{R_m}{n} \right\} \cap \left\{ p(m) - \frac{D_m}{n} \geq -20\epsilon_n \frac{R_m}{n} \right\} \right\}
\end{aligned}$$

and $\Omega_T = \Omega_\delta \cap \Omega_p$. Then there exists a constant $C > 0$ such that

$$\mathbb{P}(\Omega_\delta^c) \leq Ce^{-(\ln n)^\gamma},\ \mathbb{P}(\Omega_p^c) \leq Ce^{-\frac{1}{2}(\ln n)^\gamma},\ \mathbb{P}(\Omega_T^c) \leq Ce^{-\frac{1}{2}(\ln n)^\gamma}.$$

Proof :

Let $K > 1$ be a constant to be chosen later. We apply Lemma 6.8 in the appendix to $u = s_m - s_{m'}$, $S = S_m + S_{m'}$, $L = id$, $x = K^2 l_{n,\gamma}(R_m, R_{m'})$. For all $\eta > 0$, for all m, m' in $\mathcal{M}_n$, on an event of probability larger than $1 - e^{-K^2 l_{n,\gamma}(R_m,R_{m'})}$,

$$\delta(m,m') \leq \frac{\eta}{2}\|s_m - s_{m'}\|^2 + \frac{2v_{m,m'}^2 K^2 l_{n,\gamma}(R_m,R_{m'}) + e_{m,m'}(K^2 l_{n,\gamma}(R_m,R_{m'}))^2/9}{\eta n}. \quad (36)$$

From **[V]**, for all m, m' in $\mathcal{M}_n$,

$$2v_{m,m'}^2 K^2 l_{n,\gamma}(R_m,R_{m'})) + \frac{e_{m,m'}(K^2 l_{n,\gamma}(R_m,R_{m'}))^2}{9} \leq \left(2(K\epsilon_n)^2 + \frac{(K\epsilon_n)^4}{9} \right) \frac{R_m \vee R_{m'}}{n}.$$

Moreover, for all m, m' in $\mathcal{M}_n$,

$$\|s_m - s_{m'}\|^2 \leq 2(\|s - s_m\|^2 + \|s - s_{m'}\|^2) \leq 2(R_m + R_{m'}) \leq 4(R_m \vee R_{m'}).$$

Let $e_n(K) = \sqrt{(K\epsilon_n)^2 + (K\epsilon_n)^4/18}$. In (36) we take $\eta = e_n(K)$ and we obtain

$$\mathbb{P}\left(\delta(m,m') > 4e_n(K) \frac{R_m \vee R_{m'}}{n} \right) \leq e^{-K l_{n,\gamma}(R_m,R_{m'})}. \quad (37)$$

From (35), for all $K > 1$,

$$\mathbb{P}\left(\forall (m,m') \in \mathcal{M}_n^2,\ \delta(m,m') > 4e_n(K) \frac{R_m \vee R_{m'}}{n} \right) \leq (\Sigma(K))^2 e^{-K(\ln n)^2}.$$

Let $K = 1.1$ and take n sufficiently large so that $K^4\epsilon_n^2/18 \leq 1$, then $4e_n(K) \leq 6\epsilon_n$. Hence, the first conclusion of Lemma 5.2 holds for sufficiently large n, it holds in general, provided that we increase the constant C if necessary.

We apply Assumption **[V]** (see (32)) with $m = m'$, let $l_m = l_{n,\gamma}(R_m, R_m)$, for all $K > 0$, for all n such that $4.06(K\epsilon_n)^3 \leq 2$,

$$\frac{D_m^{3/4}(e_m(K^2 l_m)^2)^{1/4} + 0.7\sqrt{D_m v_m^2 K^2 l_m} + 0.15 v_m^2 K^2 l_m + e_m(K^2 l_m)^2}{n}$$
$$\leq (1.7K\epsilon_n + 0.15(K\epsilon_n)^2 + (K\epsilon_n)^4)\frac{R_m}{n} \leq 3K\epsilon_n \frac{R_m}{n}.$$

$$\frac{1.8D_m^{3/4}(e_m(K^2 l_m)^2)^{1/4} + 1.71\sqrt{D_m v_m^2 (K^2 l_m)} + 4.06 e_m(K^2 l_m)^2}{n}$$
$$\leq (3.51K\epsilon_n + 4.06(K\epsilon_n)^4)\frac{R_m}{n} \leq 6K\epsilon_n \frac{R_m}{n}.$$

It comes then from Proposition 2.1 applied with $x = K^2 l_m$ that, for all m in $\mathcal{M}_n$

$$\mathbb{P}\left(p(m) - \frac{D_m}{n} > 3K\epsilon_n \frac{R_m}{n}\right) \leq e^{-\frac{K^2}{20} l_m}.$$

Thus, from (34), for all $K > \sqrt{10}$, and for all n sufficiently large,

$$\mathbb{P}\left(\forall m \in \mathcal{M}_n,\ p(m) - \frac{D_m}{n} > 3K\epsilon_n \frac{R_m}{n}\right) \leq \Sigma(K^2/10) e^{-\frac{K^2}{20}(\ln n)^\gamma}.$$

We use the same arguments to prove that

$$\mathbb{P}\left(\forall m \in \mathcal{M}_n,\ p(m) - \frac{D_m}{n} < 6K\epsilon_n \frac{R_m}{n}\right) \leq \Sigma(K^2/10) e^{-\frac{K^2}{20}(\ln n)^\gamma}.$$

Fixe $K = \sqrt{10.5}$, then for all n sufficiently large , the conclusion of Lemma 5.2 holds. It holds in general provided that we increase the constant C if necessary.

Lemma 5.3 *Let $(\psi_\lambda)_{\lambda\in\Lambda}$ be an orthonormal system in $L^2(\mu)$ and let L be a linear functional defined on $L^2(\mu)$. Let $p(\Lambda) = \sum_{\lambda\in\Lambda}(\nu_n(L(\psi_\lambda)))^2$. Let $(W_1, ..., W_n)$ be a resampling scheme, let $\bar{W}_n = \sum_{i=1}^n W_i/n$ and let $v_W^2 = Var(W_1 - \bar{W}_n)$. Let*

$$D_\Lambda^W = n(v_W^2)^{-1} \sum_{\lambda\in\Lambda} \mathbb{E}^W\left((\nu_n^W(L(\psi_\lambda)))^2\right),$$

$T = \sum_{\lambda\in\Lambda}(L(\psi_\lambda) - PL(\psi_\lambda))^2$, $D = PT$ and

$$U = \frac{1}{n(n-1)} \sum_{i\neq j=1}^n \sum_{\lambda\in\Lambda} (L(\psi_\lambda)(X_i) - PL(\psi_\lambda))(L(\psi_\lambda)(X_j) - PL(\psi_\lambda)).$$

then

$$p(\Lambda) = \frac{1}{n}P_n T + \frac{n-1}{n}U,\ D_\Lambda^W = P_n T - U,\ p(\Lambda) - \frac{D_\Lambda^W}{n} = U,$$
$$\mathbb{E}(D_\Lambda^W) = D,\ D_\Lambda^W - D = \nu_n T - U.$$

Proof :

It is easy to check that

$$\begin{aligned} p(\Lambda) &= \sum_{\lambda\in\Lambda}(\frac{1}{n}\sum_{i=1}^{n} L(\psi_\lambda)(X_i) - PL(\psi_\lambda))^2 = \frac{1}{n^2}\sum_{i=1}^{n}(L(\psi_\lambda)(X_i) - PL(\psi_\lambda))^2 \\ &\quad + \frac{1}{n^2}\sum_{i\neq j=1}^{n}\sum_{\lambda\in\Lambda}(L(\psi_\lambda)(X_i) - PL(\psi_\lambda))(L(\psi_\lambda)(X_j) - PL(\psi_\lambda)) \\ &= \frac{1}{n}P_nT + \frac{n-1}{n}U. \end{aligned}$$

Recall that $\nu_n^W = P_n^W - \bar{W}_n P_n$. For all λ in Λ, since $\sum_{i=1}^n (W_i - \bar{W}_n) = 0$,

$$\begin{aligned} \nu_n^W(L(\psi_\lambda)) &= \frac{1}{n}\sum_{i=1}^{n}(W_i - \bar{W}_n)L(\psi_\lambda)(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}(W_i - \bar{W}_n)(L(\psi_\lambda)(X_i) - PL(\psi_\lambda)). \end{aligned}$$

Thus, if $E_{i,j} = \mathbb{E}\left((W_i - \bar{W}_n)(W_j - \bar{W}_n)\right)/v_W^2$,

$$\begin{aligned} D_\Lambda^W &= n(v_W^2)^{-1}\sum_{\lambda\in\Lambda}\mathbb{E}^W\left(\left(\frac{1}{n}\sum_{i=1}^{n}(W_i - \bar{W}_n)(L(\psi_\lambda)(X_i) - PL(\psi_\lambda))\right)^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{E}\left((W_i - \bar{W}_n)^2\right)}{v_W^2}(L(\psi_\lambda)(X_i) - PL(\psi_\lambda))^2 + \\ &\quad \frac{1}{n}\sum_{i\neq j=1}^{n}\sum_{\lambda\in\Lambda}E_{i,j}(L(\psi_\lambda)(X_i) - PL(\psi_\lambda))(L(\psi_\lambda)(X_j) - PL(\psi_\lambda)). \end{aligned}$$

Since the weights are exchangeable, for all $i = 1, .., n$, $\mathbb{E}((W_i - \bar{W}_n)^2) = \mathrm{Var}(W_1 - \bar{W}_n) = v_W^2$ and for all $i \neq j = 1, ..., n$,

$$v_W^2 E_{i,j} = \mathbb{E}\left((W_i - \bar{W}_n)(W_j - \bar{W}_n)\right) = \mathbb{E}\left((W_1 - \bar{W}_n)(W_2 - \bar{W}_n)\right).$$

Moreover, since $\sum_{i=1}^n (W_i - \bar{W}_n) = 0$,

$$\begin{aligned} 0 &= E\left[\left(\sum_{i=1}^{n}(W_i - \bar{W}_n)\right)^2\right] = \sum_{i=1}^{n}\mathbb{E}\left((W_i - \bar{W}_n)^2\right) + \sum_{i\neq j=1}^{n} v_W^2 E_{i,j} \\ &= n\mathbb{E}((W_1 - \bar{W}_n)^2) + n(n-1)\mathbb{E}\left((W_1 - \bar{W}_n)(W_2 - \bar{W}_n)\right). \end{aligned}$$

Hence, for all $i \neq j = 1, ..., n$, $E_{i,j} = -1/(n-1)$, thus

$$D_\Lambda^W = P_nT - U.$$

The last inequalities of Lemma 5.3 follow from the fact that $\mathbb{E}(U) = 0$. Finally,

$$p(\Lambda) - \frac{D_\Lambda^W}{n} = \frac{1}{n}P_nT + \frac{n-1}{n}U - \left(\frac{1}{n}P_nT - \frac{1}{n}U\right) = U.$$

Lemma 5.4 *Let*

$$\begin{aligned}\Omega_u &= \bigcap_{m\in\mathcal{M}_n}\left\{\frac{D_m^W}{n}-p(m)\leq 10\epsilon_n\frac{R_m}{n}\right\}\\ \Omega_l &= \bigcap_{m\in\mathcal{M}_n}\left\{\frac{D_m^W}{n}-p(m)\geq -12\epsilon_n\frac{R_m}{n}\right\}\end{aligned}$$

and $\tilde{\Omega}_p=\Omega_u\cap\Omega_l$. *There exists a constant* $C>0$ *such that* $\mathbb{P}(\tilde{\Omega}_p^c)\leq Ce^{-\frac{1}{2}(\ln n)^\gamma}$.

Proof :

From Assumption **[V]** applied with $m=m'$, (see (32)), if $l_m=l_{n,\gamma}(R_m,R_m)$, for all $K>0$,

$$D_m^{3/4}(e_m(K^2l_m)^2)^{1/4}\leq K\epsilon_nR_m,\ \sqrt{v_m^2D_m(K^2l_m)}\leq K\epsilon_nR_m,$$
$$v_m^2(K^2l_m)\leq(K\epsilon_n)^2R_m,\ e_m(Kl_m)^2\leq(K\epsilon_n)^4R_m.$$

We apply Proposition 2.4 with $x=K^2l_m$ and we obtain

$$\mathbb{P}\left(\frac{D_m^W}{n}-p(m)>\left(8.31K\epsilon_n+3(K\epsilon_n)^2+(19.1)^2(K\epsilon_n)^4\right)\frac{R_m}{n-1}\right)\leq 2e^{-K^2l_m}.$$

Thus, for all $K>1/(\sqrt{2})$, if $e_n(K)=n\left(8.31K\epsilon_n+3(K\epsilon_n)^2+(19.1)^2(K\epsilon_n)^4\right)/(n-1)$, from (34)

$$\mathbb{P}\left(\forall m\in\mathcal{M}_n,\ \frac{D_m^W}{n}-p(m)>e_n(K)\frac{R_m}{n}\right)\leq 2\Sigma(2K^2)e^{-K^2(\ln n)^\gamma}.$$

Take $K=8/8.31$ and $n\geq 10$ sufficiently large to ensure that $3K^2\epsilon_n+(19.1)^2K^4\epsilon_n^3\leq 1$, then

$$e_n(K)\leq\frac{10}{9}\left(8\epsilon_n+\epsilon_n\right)\leq 10\epsilon_n.$$

We deduce that, for sufficiently large n,

$$\mathbb{P}(\Omega_u^c)\leq 2\Sigma(2K^2)e^{-K^2(\ln n)^\gamma}.$$

We also apply Proposition 2.4 with $x=K^2l_m$, and we use the same arguments to prove that, for $K=16/16.61$, for all $n\geq 10$ sufficiently large to ensure that $(40.3)^2K^4\epsilon_n^3\leq 2$

$$\mathbb{P}\left(\forall m\in\mathcal{M}_n,\ \frac{D_m^W}{n}-p(m)<-20\epsilon_n\frac{R_m}{n}\right)\leq 3.8\Sigma(2K^2)e^{-K^2(\ln n)^\gamma}.$$

Hence, the conclusion of Lemma 5.4 holds for sufficiently large n. It holds in general, provided that we increase the constant C if necessary.

5.3 Proof of Theorem 2.2

If $c_n<0$, there is nothing to prove. We can then assume that $c_n\geq 0$, this implies in particular that

$$28\epsilon_n\leq\delta_n<1.$$

We use the notations of Lemma 5.2. From Lemma 5.2, the inequalities (19) will be proved if, on Ω_T, $D_{\hat{m}}\geq c_nD_{m^*}$ and

$$\|s-\tilde{s}\|^2\geq\frac{c_n}{5h_n^o}\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_m\|^2.$$

Let $m_o \in \arg\min_{m\in\mathcal{M}_n} R_m$, $\hat{m}$ minimizes over $\mathcal{M}_n$ the following criterion.

$$\begin{aligned}\mathrm{Crit}(m) &= P_nQ(\hat{s}_m) + \mathrm{pen}(m) + \|s\|^2 + 2\nu_n(s_{m_o}) \\ &= \|s - s_m\|^2 - p(m) + \delta(m_o, m) + \mathrm{pen}(m).\end{aligned}$$

Recall that $0 \le \mathrm{pen}(m) \le (1-\delta_n)D_m/n$. On Ω_T, for all m in $\mathcal{M}_n$, since $R_m \ge R_{m_o}$,

$$\begin{aligned}\mathrm{Crit}(m) &\ge \|s - s_m\|^2 - \frac{D_m}{n} - 16\epsilon_n\frac{R_m}{n} \ge -(1+16\epsilon_n)\frac{D_m}{n}. \\ \mathrm{Crit}(m) &\le \|s - s_m\|^2 + 26\epsilon_n\frac{R_m}{n} - \delta_n\frac{D_m}{n} = (1+26\epsilon_n)\|s - s_m\|^2 - (\delta_n - 26\epsilon_n)\frac{D_m}{n}.\end{aligned}$$

When $D_m \le c_n D_{m^*}$,

$$(1+16\epsilon_n)D_m \le D_{m^*}\left((\delta_n - 26\epsilon_n) - (1+26\epsilon_n)\frac{n\|s - s_{m^*}\|^2}{D_{m^*}}\right).$$

Thus $\mathrm{Crit}(m) \ge \mathrm{Crit}(m^*)$. This implies that $D_{\hat{m}} \ge c_n D_{m^*}$.
Moreover, on Ω_T, we also have, for all m in $\mathcal{M}_n$

$$\|s - \tilde{s}\|^2 = \frac{R_{\hat{m}}}{n} + \left(p(\hat{m}) - \frac{D_{\hat{m}}}{n}\right) \ge (1-20\epsilon_n)\frac{R_{\hat{m}}}{n},$$

and

$$\inf_{m\in\mathcal{M}_n}\|s - \hat{s}_m\|^2 \le \inf_{m\in\mathcal{M}_n}\frac{R_m}{n}(1+10\epsilon_n) \le \frac{R_{m_o}}{n}(1+10\epsilon_n).$$

Thus

$$\begin{aligned}\|s - \tilde{s}\|^2 &\ge (1-20\epsilon_n)\frac{R_{\hat{m}}}{n} \ge (1-20\epsilon_n)\frac{D_{\hat{m}}}{n} \ge (1-20\epsilon_n)c_n\frac{D_{m^*}}{n} \\ &\ge c_n\frac{1-20\epsilon_n}{h_n^o}\frac{R_{m_o}}{n} \ge \frac{c_n}{h_n^o}\frac{1-20\epsilon_n}{1+10\epsilon_n}\inf_{m\in\mathcal{M}_n}\|s - \hat{s}_m\|^2.\end{aligned}$$

We conclude the proof, saying that $\epsilon_n \le 1/28$ implies that $(1-20\epsilon_n)(1+10\epsilon_n)^{-1} \ge 8/38 \ge 1/5$.

5.4 Proof of Theorem 2.3

If $\delta_- - 46\epsilon_n < -1$, there is nothing to prove, hence, we can assume in the following that $\delta_- - 46\epsilon_n > -1$.
We keep the notation Ω_T introduced in Lemma 5.2. Let

$$\Omega_{\mathrm{pen}} = \bigcap_{m\in\mathcal{M}_n}\left\{\frac{2D_m}{n} + \delta_-\frac{R_m}{n} \le \mathrm{pen}(m) \le \frac{2D_m}{n} + \delta^+\frac{R_m}{n}\right\},$$

$\Omega = \Omega_T \cap \Omega_{\mathrm{pen}}$ and $m_o \in \arg\min_{m\in\mathcal{M}_n} R_m$. Recall that $\mathbb{P}(\Omega_{\mathrm{pen}}) \ge 1 - p'$ and that, $\hat{m}$ minimizes over m the following criterion.

$$\begin{aligned}\mathrm{Crit}(m) &= P_nQ(\hat{s}_m) + \mathrm{pen}(m) + \|s\|^2 + 2\nu_n(s_{m_o}) \\ &= \|s - s_m\|^2 - p(m) + \delta(m_o, m) + \mathrm{pen}(m).\end{aligned}$$

Therefore, on Ω, for all m in $\mathcal{M}_n$, since $R_m \geq R_{m_o}$,

$$\begin{aligned}
\mathrm{Crit}(m) &\geq (1+\delta_-)\frac{R_m}{n} + \left(\frac{D_m}{n} - p(m)\right) - 6\epsilon_n \frac{R_m}{n} \\
&\geq (1+\delta_- - 16\epsilon_n)\|s-s_m\|^2 + (1+\delta_- - 16\epsilon_n)\frac{D_m}{n} \geq (1+\delta_- - 16\epsilon_n)\frac{D_m}{n} \\
\mathrm{Crit}(m) &\leq (1+\delta^+ + 26\epsilon_n)\frac{R_m}{n}.
\end{aligned}$$

If $D_m > C_n(\delta_-, \delta^+)R_{m_o}$,

$$(1+\delta_- - 16\epsilon_n)D_m > (1+\delta^+ + 26\epsilon_n)R_{m_o},$$

Thus $\mathrm{Crit}(m) > \mathrm{Crit}(m_o)$, hence $D_{\hat{m}} \leq C_n(\delta_-, \delta^+)R_{m_o}$.
Moreover, from (6), for all m in $\mathcal{M}_n$

$$\begin{aligned}
\|s-\tilde{s}\|^2 &\leq \|s-\hat{s}_m\|^2 + (\mathrm{pen}(m) - 2p(m)) + (2p(\hat{m}) - \mathrm{pen}(\hat{m})) + \delta(\hat{m}, m) \\
&\leq \|s-\hat{s}_m\|^2 + 2\left(\frac{D_m}{n} - p(m)\right) + (\delta^+ + 6\epsilon_n)\frac{R_m}{n} \\
&\quad + 2\left(p(\hat{m}) - \frac{D_{\hat{m}}}{n}\right) + (-\delta_- + 6\epsilon_n)\frac{R_{\hat{m}}}{n} \\
&\leq \|s-\hat{s}_m\|^2 + (46\epsilon_n + \delta^+)\frac{R_m}{n} + (26\epsilon_n - \delta_-)\frac{R_{\hat{m}}}{n}.
\end{aligned}$$

For all m in $\mathcal{M}_n$, on Ω_T,

$$\|s-\hat{s}_m\|^2 = \frac{R_m}{n} + \left(p(m) - \frac{D_m}{n}\right) \geq (1-20\epsilon_n)\frac{R_m}{n}.$$

Hence, for all $m \in \mathcal{M}_n$,

$$\|s-\tilde{s}\|^2 \leq \|s-\hat{s}_m\|^2 \left(1 + \frac{46\epsilon_n + \delta^+}{1-20\epsilon_n}\right) + \frac{26\epsilon_n - \delta_-}{1-20\epsilon_n}\|s-\tilde{s}\|^2.$$

This concludes the proof of Proposition 2.3.

5.5 Proof of Proposition 2.4

We apply Lemma 5.3 with $L = id$ and $\Lambda = m$. By definition of $p(m)$ and D_m^W,

$$p(m) - \frac{D_m^W}{n} = \frac{1}{n(n-1)} \sum_{i \neq j=1}^{n} \sum_{\lambda \in m} (\psi_\lambda(X_i) - P\psi_\lambda)(\psi_\lambda(X_j) - P\psi_\lambda).$$

Thus, from Lemma 6.7 in the appendix, for all $x > 0$,

$$\mathbb{P}\left(p(m) - \frac{D_m^W}{n} > \frac{5.31 D_m^{3/4}(e_m x^2)^{1/4} + 3\sqrt{v_m^2 D_m x} + 3v_m^2 x + e_m(19.1x)^2}{n-1}\right) \leq 2e^{-x}.$$

$$\mathbb{P}\left(\frac{D_m^W}{n} - p(m) > \frac{9 D_m^{3/4}(e_m x^2)^{1/4} + 7.61\sqrt{v_m^2 D_m x} + e_m(40.3x)^2}{n-1}\right) \leq 3.8e^{-x}.$$

This proves (23) and (24).
In order to obtain (21) and (22), we introduce, for all m in $\mathcal{M}_n$, the function $T_m = \sum_{\lambda\in m}(\psi_\lambda - P\psi_\lambda)^2$ and the random variable

$$U_m = \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n}\sum_{\lambda\in m}(\psi_\lambda(X_i)-P\psi_\lambda)(\psi_\lambda(X_j)-P\psi_\lambda).$$

We apply Lemma 5.3 with $L = id$, we obtain

$$D_m^W - D_m = \nu_n(T_m) - U_m.$$

From Bernstein's inequality (see Proposition 6.3), for all $x > 0$ and all ξ in $\{-1, 1\}$,

$$\mathbb{P}\left(\xi\nu_n(T_m) > \sqrt{\frac{2\mathrm{Var}(T_m(X))x}{n}} + \frac{\|T_m\|_\infty x}{3n}\right) \leq e^{-x}.$$

From Cauchy-Schwarz inequality, $T_m = \sup_{t\in B_m}(t-Pt)^2$, thus $\|T_m\|_\infty/n = 4e_m$ and $\mathrm{Var}(T_m(X))/n \leq \|T_m\|_\infty PT_m/n = 4e_m D_m$, therefore, for all $x > 0$ and all ξ in $\{-1, 1\}$,

$$\mathbb{P}\left(\xi\nu_n(T_m) > \sqrt{8e_m D_m x} + \frac{4e_m x}{3}\right) \leq e^{-x}.$$

Moreover, from Lemma 6.7 in the appendix, for all $x > 0$,

$$\mathbb{P}\left(U_m > \frac{5.31D_m^{3/4}(e_m x^2)^{1/4} + 3\sqrt{v_m^2 D_m x} + 3v_m^2 x + e_m(19.1x)^2}{n-1}\right) \leq 2e^{-x}.$$

$$\mathbb{P}\left(U_m < -\frac{9D_m^{3/4}(e_m x^2)^{1/4} + 7.61\sqrt{v_m^2 D_m x} + e_m(40.3x)^2}{n-1}\right) \leq 3.8e^{-x}.$$

We deduce that, for all $x > 0$, with probability larger than $1 - 4.8e^{-x}$,

$$\begin{aligned} D_m^W - D_m &\leq \sqrt{8e_m D_m x} + e_m\left(\frac{4x}{3} + \frac{(40.3x)^2}{n-1}\right) \\ &\quad + \frac{9D_m^{3/4}(e_m x^2)^{1/4} + 7.61\sqrt{v_m^2 D_m x}}{n-1}. \end{aligned}$$

Moreover, for all $x > 0$, on an event of probability larger than $1 - 3e^{-x}$,

$$\begin{aligned} D_m^W - D_m &\geq -\sqrt{8e_m D_m x} - e_m\left(\frac{4x}{3} + \frac{(19.1x)^2}{n-1}\right) \\ &\quad - \frac{5.31D_m^{3/4}(e_m x^2)^{1/4} + 3\sqrt{v_m^2 D_m x} + 3v_m^2 x}{n-1}. \end{aligned}$$

5.6 Proof of Theorem 2.5

Recall that $\mathbb{P}(\Omega_T^c) \leq Ce^{-\frac{1}{2}(\ln n)^\gamma}$, and that, on Ω_T,

$$\forall m \in \mathcal{M}_n, (1 - 20\epsilon_n)\frac{R_m}{n} \leq \|s - \hat{s}_m\|^2,$$

$$\forall m, m' \in \mathcal{M}_n^2,\ \delta(m, m') \leq 6\epsilon_n \frac{R_m \vee R_{m'}}{n}.$$

Let $\tilde{\Omega}_p$ be the event defined in Lemma 5.4 and let $\Omega = \tilde{\Omega}_p \cap \Omega_T$, from Lemma 5.2, $\mathbb{P}(\Omega^c) \leq Ce^{-\frac{1}{2}(\ln n)^\gamma}$. Recall that $\text{pen}(m) = 2D_m^W/n$. On Ω, from (6), for all n such that $20\epsilon_n < 1$, for all m in $\mathcal{M}_n$,

$$\begin{aligned}\|s - \tilde{s}\|^2 &\leq \|s - \hat{s}_m\|^2 + 26\epsilon_n \frac{R_m}{n} + 16\epsilon_n \frac{R_{\hat{m}}}{n} \\ &\leq \|s - \hat{s}_m\|^2 + \frac{26\epsilon_n}{1 - 20\epsilon_n}\|s - \hat{s}_m\|^2 + \frac{16\epsilon_n}{1 - 20\epsilon_n}\|s - \tilde{s}\|^2.\end{aligned}$$

Hence, for all n such that $20\epsilon_n < 1$, on Ω,

$$(1 - 36\epsilon_n)\|s - \tilde{s}\|^2 \leq (1 + 6\epsilon_n) \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2.$$

For all n such that $42/(1 - 36\epsilon_n) < 100$,

$$\|s - \tilde{s}\|^2 \leq \left(1 + \frac{42\epsilon_n}{1 - 36\epsilon_n}\right) \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2 \leq (1 + 100\epsilon_n) \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2.$$

Hence (25) holds for sufficiently large n, it holds in general provided that we enlarge the constant C if necessary..

6 Appendix

In this Section, we state and prove some technical lemmas that are useful in the proofs. The main tool is the first Lemma based on Bousquet's version of Talagrand's inequality. It is a concentration inequality for the square of the supremum of the empirical process over a uniformly bounded class of functions. Recall first Bousquet's [10] and Klein & Rio [17] versions of Talagrand's inequality.

Theorem 6.1 *(Bousquet's bound) Let $X_1, ..., X_n$ be i.i.d. random variables valued in a measurable space $(\mathbb{X}, \mathcal{X})$ and let S be a class of real valued functions bounded by b. Let $v^2 = \sup_{t \in S} Var(t(X))$ and let $Z = \sup_{t \in S} \nu_n t$. Then*

$$\forall x > 0,\ \mathbb{P}\left(Z > \mathbb{E}(Z) + \sqrt{\frac{2}{n}(v^2 + 2b\mathbb{E}(Z))x} + \frac{bx}{3n}\right) \leq e^{-x}.$$

Theorem 6.2 *(Klein & Rio's bound) Let $X_1, ..., X_n$ be i.i.d. random variables valued in a measurable space $(\mathbb{X}, \mathcal{X})$ and let S be a class of real valued functions bounded by b. Let $v^2 = \sup_{t \in S} Var(t(X))$ and let $Z = \sup_{t \in S} \nu_n t$. Then*

$$\forall x > 0,\ \mathbb{P}\left(Z < \mathbb{E}(Z) - \sqrt{\frac{2}{n}(v^2 + 2b\mathbb{E}(Z))x} - \frac{8bx}{3n}\right) \leq e^{-x}.$$

Let us now also recall Bernstein's inequality.

Proposition 6.3 *Bernstein's inequality*
Let $X_1, ..., X_n$ be iid random variables valued in a measurable space $(X, \mathcal{X})$ and let t be a measurable real valued function. Then, for all $x > 0$,

$$\mathbb{P}\left(\nu_n(t) > \sqrt{\frac{2\,Var(t(X_1))x}{n}} + \frac{\|t\|_\infty\, x}{3n}\right) \leq e^{-x}.$$

We derive from these bounds the following useful corollary. Hereafter, S denotes a symetric class of real valued functions upper bounded by b, $v^2 = \sup_{t\in S} \mathrm{Var}(t(X))$, $Z = \sup_{t\in S} \nu_n t$, $n\mathbb{E}(Z^2) = D$. Since S is symetric, we always have $Z \geq 0$.

Corollary 6.4 *Let S be a symetric class of real valued functions upper bounded by b, $v^2 = \sup_{t\in S} Var(t(X))$, $Z = \sup_{t\in S} \nu_n t$, $n\mathbb{E}(Z^2) = D$, $e_b = b^2/n$ and*

$$nE_m = 225e_b + \left(2.1 + \sqrt{2\pi}\right)\sqrt{v^2 D} + \sqrt{15}D^{3/4}e_b^{1/4},$$

then

$$\mathbb{E}(Z^2\mathbf{1}_{Z\geq\mathbb{E}(Z)}) \leq (\mathbb{E}(Z))^2\mathbb{P}\left(Z \geq \mathbb{E}(Z)\right) + E_m. \tag{38}$$

In particular,

$$(\mathbb{E}(Z))^2 \leq \mathbb{E}(Z^2) \leq (\mathbb{E}(Z))^2 + E_m. \tag{39}$$

Proof :

We have

$$\begin{aligned}\mathbb{E}(Z^2\mathbf{1}_{Z\geq\mathbb{E}(Z)}) &= \int_0^\infty \mathbb{P}(Z^2\mathbf{1}_{Z\geq\mathbb{E}(Z)} > x)dx = \int_0^\infty \mathbb{P}(Z\mathbf{1}_{Z\geq\mathbb{E}(Z)} > \sqrt{x})dx \\ &= (\mathbb{E}(Z))^2\mathbb{P}\left(Z \geq \mathbb{E}(Z)\right) + \int_{(\mathbb{E}(Z))^2}^\infty \mathbb{P}(Z > \sqrt{x})dx\end{aligned}$$

Take $x = (\mathbb{E}(Z) + \sqrt{2(v^2 + 2b\mathbb{E}(Z))y/n} + by/(3n))^2$ in the previous integral, from Bousquet's version of Talagrand's inequality,

$$\begin{aligned}\mathbb{E}(Z^2\mathbf{1}_{Z\geq\mathbb{E}(Z)}) &\leq \mathbb{E}(Z)\sqrt{\frac{2}{n}(v^2 + 2b\mathbb{E}(Z))}\int_0^\infty \frac{e^{-y}}{\sqrt{y}}dy + \frac{2v^2 + 14b\mathbb{E}(Z)/3}{n}\int_0^\infty e^{-y}dy \\ &\quad + \frac{b}{n}\sqrt{\frac{2}{n}(v^2 + 2b\mathbb{E}(Z))}\int_0^\infty e^{-y}\sqrt{y}dy + \frac{2b^2}{9n^2}\int_0^\infty ye^{-y}dy.\end{aligned}$$

Classical computations lead to

$$\int_0^\infty \frac{e^{-y}}{\sqrt{y}}dy = 2\int_0^\infty e^{-y}\sqrt{y}dy = \sqrt{\pi},\ \int_0^\infty e^{-y}dy = \int_0^\infty ye^{-y}dy = 1.$$

Therefore, if $e_b = b^2/n$, using repeatedly the inequalities

$$a^\alpha b^{1-\alpha} \leq \alpha a + (1-\alpha)b \tag{40}$$

and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we obtain, for all $\eta > 0$,

$$\sqrt{ne_b}\mathbb{E}(Z) \leq \frac{e_b}{3\eta^2} + \frac{2\eta}{3}e_b^{1/4}(\sqrt{n}\mathbb{E}(Z))^{3/2},$$

$$(\sqrt{n}\mathbb{E}(Z))^{1/2}e_b^{3/4} \leq \frac{\eta}{3}e_b^{1/4}(\sqrt{n}\mathbb{E}(Z))^{3/2} + \frac{2e_b}{3\sqrt{\eta}}.$$

Thus

$$\begin{aligned}\mathbb{E}(Z^2\mathbf{1}_{Z\geq\mathbb{E}(Z)}) &\leq \left(2v^2 + \frac{2}{9}e_b + v\frac{\sqrt{2\pi e_b}}{2}\right)\frac{1}{n} + \sqrt{\pi}\frac{\sqrt{\sqrt{n}\mathbb{E}(Z)}\,(e_b)^{3/4}}{n} \\ &\quad + \left(\frac{14}{3}\sqrt{e_b} + v\sqrt{2\pi}\right)\frac{\sqrt{n}\mathbb{E}(Z)}{n} + 2\sqrt{\pi}\frac{(\sqrt{n}\mathbb{E}(Z))^{3/2}\,(e_b)^{1/4}}{n} \\ &\leq \left(2 + \eta\frac{\sqrt{2\pi}}{4}\right)\frac{v^2}{n} + \sqrt{\frac{2\pi}{n}}v\mathbb{E}(Z) + \left(\frac{2}{9} + \frac{\sqrt{2\pi}}{4\eta} + \frac{2\sqrt{\pi}}{3\sqrt{\eta}} + \frac{14}{9\eta^2}\right)\frac{e_b}{n} \\ &\quad + \left(\eta\left(\frac{\sqrt{\pi}}{3} + \frac{28}{9}\right) + 2\sqrt{\pi}\right)\frac{(\sqrt{n}\mathbb{E}(Z))^{3/2}\,(e_b)^{1/4}}{n}.\end{aligned}$$

Therefore, taking $\eta = 0.088$, we obtain

$$\mathbb{E}(Z^2 \mathbf{1}_{Z \geq \mathbb{E}(Z)}) \leq 2.1 \frac{v^2}{n} + 15^2 \frac{e_b}{n} + \sqrt{2\pi} v \frac{\sqrt{n}\mathbb{E}(Z)}{n} + \sqrt{15} \frac{(\sqrt{n}\mathbb{E}(Z))^{3/2} (e_b)^{1/4}}{n}.$$

Finally, we use Cauchy-Schwarz inequality to obtain that $\sqrt{n}\mathbb{E}(Z) \leq (n\mathbb{E}(Z^2))^{1/2} = (D)^{1/2}$. Since $v^2 \leq D$, we get (38).

We deduce from this result the following concentration inequalities for Z^2

Corollary 6.5 *Let $e_b = b^2/n$. We have, for all $x > 0$,*

$$\mathbb{P}\left(Z^2 - \frac{D}{n} > \frac{D^{3/4}(e_b(19x)^2)^{1/4} + 3\sqrt{Dv^2x}) + 3v^2x + e_b(19x)^2}{n} \right) \leq e^{-x}.$$

Moreover, for all $x > 0$, with probability larger than $1 - e^{-x}$,

$$\frac{D}{n} - Z^2 \leq \frac{D^{3/4} e_b^{1/4}(\sqrt{15} + 4.127\sqrt{x}) + \sqrt{v^2 D}(4.61 + 3\sqrt{x}) + 225 e_b(6.2x^2 + 1)}{n}. \quad (41)$$

Proof :

From Bousquet's version of Talagrand's inequality and from $(\mathbb{E}(Z))^2 \leq \mathbb{E}(Z^2)$, we obtain that, for all $x > 0$, with probability larger than $1 - e^{-x}$, $Z^2 - D/n$ is not larger than

$$\frac{4D^{3/4}(e_b x^2)^{1/4} + \sqrt{D}(14\sqrt{e_b x^2}/3 + 2\sqrt{2v^2x}) + 4D^{1/4}(e_b x^2)^{3/4}/3 + 3v^2x + e_b x^2/3}{n}.$$

We use repeatedly the inequality $a^\alpha b^{1-\alpha} \leq \alpha a + (1-\alpha)b$ to obtain that, with probability at least $1 - e^{-x}$, $Z^2 - D/n$ is not larger than

$$\frac{(4 + 32\eta/9)D^{3/4}(e_b x^2)^{1/4} + 2\sqrt{2}\sqrt{Dv^2x} + 3v^2x + (3 + 14/\eta^2 + 8/\sqrt{\eta}) e_b x^2/9}{n}.$$

For $\eta = 0.07$, this gives

$$Z^2 - \frac{D}{n} > \frac{D^{3/4}(e_b(19x)^2)^{1/4} + 2\sqrt{2}\sqrt{Dv^2x} + 3v^2x + e_b(19x)^2}{n}.$$

For the second one we use Klein's version of Talagrand's inequality to obtain, for all $x > 0$ such that $r(x) = \sqrt{2(v^2 + 2b\mathbb{E}(Z))x/n} + 8bx/3n < \mathbb{E}(Z)$,

$$\mathbb{P}\left(Z^2 < (\mathbb{E}(Z) - r(x))^2 \right) \leq e^{-x}.$$

We have $(\mathbb{E}(Z) - r(x))^2 = (\mathbb{E}(Z))^2 - 2\mathbb{E}(Z)r(x) + r(x)^2 \geq (\mathbb{E}(Z))^2 - 2\mathbb{E}(Z)r(x)$, thus

$$\mathbb{P}\left(Z^2 < (\mathbb{E}(Z))^2 - 2\mathbb{E}(Z)r(x) \right) \leq e^{-x}.$$

From the previous corollary, $(\mathbb{E}(Z))^2 \geq \mathbb{E}(Z^2) - E_m$, thus

$$\mathbb{P}\left(Z^2 < \mathbb{E}(Z^2) - E_m - 2\mathbb{E}(Z)r(x) \right) \leq e^{-x}.$$

In order to conclude the proof of 6.5, just remark that

$$\begin{aligned} 2\mathbb{E}(Z)r(x) &\leq \frac{4D^{3/4}(e_b x^2)^{1/4} + 3\sqrt{Dv^2x} + 16\sqrt{De_b x^2}/3}{n} \\ &\leq \frac{(4 + 32\eta/9)D^{3/4}(e_b x^2)^{1/4} + 3\sqrt{Dv^2x} + 16/(9\eta^2) e_b x^2}{n}. \end{aligned}$$

For $\eta = 0,0357$, we obtain (41).

Finally, we have obtained the following result for the concentration of Z^2 around its mean

Corollary 6.6 *For all $x > 0$,*

$$\mathbb{P}\left(Z^2 - \frac{D}{n} > \frac{D^{3/4}(e_b(19x)^2)^{1/4} + 3\sqrt{Dv^2x} + 3v^2x + e_b(19x)^2}{n}\right) \leq e^{-x}.$$

$$\mathbb{P}\left(Z^2 - \frac{D}{n} < -\frac{8D^{3/4}(e_b x^2)^{1/4} + 7.61\sqrt{v^2Dx} + e_b(40.25x)^2}{n}\right) \leq ee^{-x}.$$

Proof :

In order to obtain the second inequality, we remark that the inequality is trivial when $x \leq 1$, thus we only have to use (41) for $x > 1$ and then $\sqrt{x} > 1$ and $x^2 > 1$.

We will use this lemma to obtain a concentration inequality for totally degenerate U-statistics of order 2. The following result generalizes a previous inequality due to Houdré & Reynaud-Bouret [16] to random variables taking values in a measurable space.

Lemma 6.7 *Let $X, X_1, ..., X_n$ be i.i.d random variables taking value in a measurable space $(\mathbb{X}, \mathcal{X})$ with common law P. Let μ be a measure on $(\mathbb{X}, \mathcal{X})$ and let $(t_\lambda)_{\lambda\in\Lambda}$ be a set of functions in $L^2(\mu)$. Let*

$$B = \{t = \sum_{\lambda\in\Lambda} a_\lambda t_\lambda,\ \sum_{\lambda\in\Lambda} a_\lambda^2 \leq 1\},\ D = \mathbb{E}\left(\sup_{t\in B}(t(X) - Pt)^2\right),$$

$$v^2 = \sup_{t\in B} Var(t(X)),\ b = \sup_{t\in B} \|t\|_\infty \text{ and } e_b = \frac{b^2}{n}.$$

Let

$$U = \frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} \sum_{\lambda\in\Lambda} (t_\lambda(X_i) - Pt_\lambda)(t_\lambda(X_j) - Pt_\lambda).$$

Then the following inequality holds

$$\forall x > 0,\ \mathbb{P}\left(U > \frac{5.31D^{3/4}(e_b x^2)^{1/4} + 3\sqrt{v^2Dx} + 3v^2x + e_b(19.1x)^2}{n-1}\right) \leq 2e^{-x}. \quad (42)$$

$$\forall x > 0,\ \mathbb{P}\left(U < -\frac{9D^{3/4}(e_b x^2)^{1/4} + 7.61\sqrt{v^2Dx} + e_b(40.3x)^2}{n-1}\right) \leq 3.8e^{-x}. \quad (43)$$

Proof :

Remark that, from Cauchy-Schwarz inequality,

$$\sup_{t\in B}(\nu_n(t))^2 = \left(\sup_{\sum a_\lambda^2 \leq 1} \sum_{\lambda\in\Lambda} a_\lambda \nu_n(t_\lambda)\right)^2 = \sum_{\lambda\in\Lambda} (\nu_n(t_\lambda))^2.$$

For all x in $\mathbb{X}$, from Cauchy-Schwarz inequality,

$$\sup_{t\in B}(t(x) - Pt)^2 = \sum_{\lambda} (t_\lambda(x) - Pt_\lambda)^2,$$

in particular, $D = \sum_{\lambda\in\Lambda} \mathrm{Var}(\psi_\lambda(X))$. Moreover, easy algebra leads to

$$\begin{aligned}\sum_{\lambda\in\Lambda}(\nu_n(t_\lambda))^2 &= \frac{1}{n^2}\sum_{i=1}^n\sum_{\lambda\in\Lambda}(t_\lambda(X_i)-Pt_\lambda)^2 \\ &\quad +\frac{1}{n^2}\sum_{i\neq j=1}^n\sum_{\lambda\in\Lambda}(t_\lambda(X_i)-Pt_\lambda)(t_\lambda(X_j)-Pt_\lambda) \\ &= \frac{1}{n}P_n\left(\sum_{\lambda\in\Lambda}(t_\lambda-Pt_\lambda)^2\right)+\frac{n-1}{n}U.\end{aligned}$$

Let $Z^2 = \sup_{t\in B}(\nu_n(t))^2$, $T_\Lambda = \sum_{\lambda\in\Lambda}(t_\lambda - Pt_\lambda)^2$,

$$\mathbb{E}(Z^2) = \mathbb{E}\left(\frac{1}{n}P_nT_\Lambda\right) = \frac{D}{n}.$$

Hence

$$U = \frac{n}{n-1}\left(Z^2 - \mathbb{E}(Z^2) - \frac{1}{n}\nu_n(T_\Lambda)\right).$$

From Corollary 6.6, for all $x > 0$,

$$\mathbb{P}\left(Z^2 - \frac{D}{n} > \frac{D^{3/4}(e_b(19x)^2)^{1/4} + 3\sqrt{v^2Dx} + 3v^2x + e_b(19x)^2}{n}\right) \leq e^{-x}.$$

$$\mathbb{P}\left(Z^2 - \frac{D}{n} < -\frac{8D^{3/4}(e_b(x)^2)^{1/4} + 7.61\sqrt{v^2Dx} + e_b(40.25x)^2}{n}\right) \leq 2.8e^{-x}.$$

Moreover, from Bernstein inequality, for all $x > 0$,

$$\mathbb{P}\left(-\nu_nT_\Lambda > \sqrt{2De_bx} + \frac{e_bx}{3}\right) \leq e^{-x}.$$

$$\mathbb{P}\left(\nu_nT_\Lambda > \sqrt{2De_bx} + \frac{e_bx}{3}\right) \leq e^{-x}.$$

We apply inequality (40) with $a = D^{3/4}(e_bx^2)^{1/4}$, $b = e_b\sqrt{x}$, $\alpha = 2/3$ and we obtain

$$\mathbb{P}\left(-\nu_nT_\Lambda > \frac{2\sqrt{2}}{3}D^{3/4}(e_bx^2)^{1/4} + e_b\left(\frac{x+\sqrt{2x}}{3}\right)\right) \leq e^{-x}.$$

$$\mathbb{P}\left(\nu_nT_\Lambda > \frac{2\sqrt{2}}{3}D^{3/4}(e_bx^2)^{1/4} + e_b\left(\frac{x+\sqrt{2x}}{3}\right)\right) \leq e^{-x}.$$

Therefore, for all $x > 0$,

$$\mathbb{P}\left(U > \frac{5.31D^{3/4}(e_bx^2)^{1/4} + 3\sqrt{v^2Dx} + 3v^2x + e_b\left((19x)^2 + (x+\sqrt{2x})/3\right)}{n-1}\right) \leq 2e^{-x}.$$

$$\mathbb{P}\left(U < -\frac{9D^{3/4}(e_bx^2)^{1/4} + 7.61\sqrt{v^2Dx} + e_b\left((40.25x)^2 + (x+\sqrt{2x})/3\right)}{n-1}\right) \leq 3.8e^{-x}.$$

These inequalities are trivial when $x < 1$. We only use them when $x > 1$ and we obtain (42) and (43) since $x < x^2$ and $\sqrt{x} < x^2$ when $x > 1$.

Let us now state the corollary of Bernstein's inequality that we used repeatedly in the article.

Lemma 6.8 *Let $X, X_1, ..., X_n$ be i.i.d random variables taking value in a measurable space $(\mathbb{X}, \mathcal{X})$ with common law P. Let μ be a measure on $(\mathbb{X}, \mathcal{X})$ and let $(\psi_\lambda)_{\lambda\in\Lambda}$ be an orthonormal system in $L^2(\mu)$. Let L be a linear functional in $L^2(\mu)$ and let $B = \{t = \sum_{\lambda\in\Lambda} a_\lambda L(\psi_\lambda),\ \sum_{\lambda\in\Lambda} a_\lambda^2 \leq 1\}$, $v^2 = \sup_{t\in B} Var(t(X))$, $b = \sup_{t\in B} \|t\|_\infty$ and $e_b = b^2/n$. Let u be a function in S, the linear space spanned by the functions $(\psi_\lambda)_{\lambda\in\Lambda}$ and let $\eta > 0$. Then the following inequality holds*

$$\forall x > 0,\ \mathbb{P}\left(\nu_n(L(u)) > \frac{\eta}{2}\|u\|^2 + \frac{2v^2x + e_b x^2/9}{\eta n}\right) \leq e^{-x}. \tag{44}$$

Proof :

From Bernstein's inequality,

$$\forall x > 0,\ \mathbb{P}\left(\nu_n(L(u)) > \sqrt{\frac{2\mathrm{Var}(L(u)(X))x}{n}} + \frac{\|L(u)\|_\infty x}{3n}\right) \leq e^{-x}.$$

Since $t = L(u/\|u\|)$ belongs to B,

$$\begin{aligned}\sqrt{\frac{2\mathrm{Var}(L(u)(X))x}{n}} + \frac{\|L(u)\|_\infty x}{3n} &= \|u\|\left(\sqrt{\frac{2\mathrm{Var}(t(X))x}{n}} + \frac{\|t\|_\infty x}{3n}\right)\\ &\leq \frac{\eta}{2}\|u\|^2 + \frac{1}{2\eta}\left(\sqrt{\frac{2v^2x}{n}} + \frac{bx}{3n}\right)^2.\end{aligned}$$

We conclude the proof using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$.

References

[1] H. Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217, 1970.

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[3] S. Arlot. *Resampling and model selection.* PhD thesis, Université Paris-Sud 11, 2007.

[4] S. Arlot. Model selection by resampling penalization. *Electron. J. Statist.*, 3:557–624, 2009.

[5] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine learning research*, 10:245–279, 2009.

[6] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[7] L. Birgé. Model selection for density estimation with l^2-loss. *Preprint*, 2008.

[8] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[9] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[10] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[11] A. Célisse. Density estimation via cross validation: Model selection point of view. *Preprint, downloadable on arXiv.org : 08110802*, 2008.

[12] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996.

[13] B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.

[14] B. Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331, 1983.

[15] M. Fromont. Model selection by bootstrap penalization for classification. *Machine Learning*, 66(2, 3):165–207, 2007.

[16] C. Houdré and P. Reynaud-Bouret. Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 55–69. Birkhäuser, Basel, 2003.

[17] T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077, 2005.

[18] C.L. Mallows. Some comments on c_p. *Technometrics*, 15:661–675, 1973.

[19] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[20] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.*, 9(2):65–78, 1982.

[21] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

[22] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996.